NPS55-82-005

NAVAL POSTGRADUATE SCHOOL

Monterey, California

STOCHASTIC MODELING: IDEAS AND TECHNIQUES

by

Donald P. Gaver

January 1982

Approved for public release; distribution unlimited.

Prepared for:

Chief of Naval Research
ington, Virginia 22217

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral J. J. Ekelund
Superintendent

David A. Schrady
Acting Provost

Work on this report was sponsored in part by the Office of Naval Research, Arlington, VA.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55-82-005	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) STOCHASTIC MODELING: IDEAS AND TECHNIQUES		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Donald P. Gaver		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940 University Libre de Bruxelles Brussels, Belgium		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N; RR14-05-0E N000 1482WR20017
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		12. REPORT DATE January 1982 13. NUMBER OF PAGES 82
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Chief of Naval Research Arlington, Virginia 22217		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) applied probability probability modeling reliability availability waiting lines		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This report summarizes the contents of lectures given on probability modeling and reports some new results on the availability of inspected systems of redundant systems in random environments, and on "sculptured distributions".		

Stochastic Modeling: Ideas and Techniques

Donald P. Gaver

1. Introduction

The primary purpose of this chapter is to summarize the contents of lectures on stochastic modeling presented at the Université Libre de Bruxelles (ULB) in the period March-May, 1981. Much of the material selected for presentation was from the standard menu of probabilistic topics typical of a second course as given to engineers, operations researchers, statisticians, or computer scientists. An attempt was made to emphasize a modeling attitude rather than details of mathematical rigor, illustrating with problems and techniques that are not often prominent in such courses. For example, attention was given to problems of, and models for, redundant system reliability and availability, queueing with priorities, first-passage times and areas under path functions of stochastic processes, (total waiting times), and various other topics. Also included was a brief account of aspects of modern data analysis, with the implication that its usefulness is significant at the pre-modeling and model-assessment stage of an investigation.

A secondary, but gratifying, purpose is to briefly report on cooperative work initiated with faculty and students at ULB. I wish to mention the enjoyable collaboration with Dr. Guy Latouche on development of efficient computational methods for repairman-like Markov models in random environments, and with Ph. Collard on the application of sculptured distributions in the simulation

evaluation of certain scheduling algorithms. The interest and warm hospitality of Prof. Guy Louchard, head of the Dept. of Computer Science at ULB, was also much appreciated.

2. The Total Modeling Process: Brief Overview

It is coming to be recognized that the topic of mathematical modeling (including stochastic modeling) exists in its own right as a subject suitable for a formal university course; see Bender [1978]. The modeling step is part of a process of several stages or steps; these may be expressed as follows (Gaver and Thompson [1973]):

(a) Identify the general problem area or situation; identify specific questions concerning that area.

(b) Obtain and analyze subject-matter information and data relating to the problem area. Often an examination of such information and data will suggest suitably formulated questions, as in (a).

(c) Construct a preliminary model, or models, representing the important features of the situation. Deduce some model implications.

(d) Refer the result of (c) to subject-matter specialists and decision-makers for qualitative critique; revise the model accordingly. This likely means re-doing (a) - (d).

(e) Assess the empirical validity of the model to the degree possible. Check the sensitivity of model conclusions to changes in model assumptions (sub-model inputs), and to data variations. Submit to judgement by subject-matter experts -- but anticipate differences of opinion! The modeling, and re-modeling, process may help to reconcile such differences.

(f) Compute required answers to interesting questions. Assess the degree of uncertainty in these answers possibly resulting from model mis-specification, data bias or other deficiency, computational error, and sampling error in estimates of basic parameters or in simulation results used to supply model implications.

(g) Communicate, and aid in implementing, the results of the model.

(h) Monitor the situation for possible changes in the environment, and hence for the necessity to change the model.

Of course the emphasis in these notes (and in the lectures, was) upon the actual modeling step, (c). However, some attention was given to the display of data for pre-model examination (Tukey's exploratory data analysis), and to model parameter estimation techniques, particularly those robust methods that attempt to deal with questions of data deficiencies.

3. Topics in Outline

In this section we out-line the basic contents of the lectures. These were in general arranged so as to first present mathematical definitions and properties, and then to illustrate in terms of sample models for various situations.

(1) Review of Probabilistic Concepts, Particularly Conditioning.

In this lecture the following basic notions of probability were defined or reviewed: random experiment or trial, sample or event space, events and combinations of events, probability as a function with rules for combination, conditional probability, independence, and Bayes' Theorem, random variables and their moments or expectations, transforms (characteristic function, Laplace transform, and generating function) and their moment-generating, convolution, unicity, and continuity properties, plus properties of conditional expectations. In addition, certain classical univariate distributions were reviewed (Normal/Gaussian, log-Normal, exponential and gamma, etc.)

By way of illustration, a simple problem of equipment (or possibly software) unrealiability was considered.

Situation: Suppose a system is made up of components that individually fail after a time because of the action of faults; the latter may be the result of component mis-design, or attributable to bad installation or adjustment ("human error"), or to a mistake in computer program coding. We wish to relate system failure rate to initial fault content.

Model: N is a random variable (rv) representing the number of faults initially installed unwittingly in the system. Let $\{T_i, i = 1,2, \ldots, N\}$ be the sequence of rv describing

the failure time of each fault, measured from time at which the system begins use; T_i may actually be the time at which the service of the particular component is first requested. Suppose the system fails at

$$X_N = \min \{T_1, T_2, \ldots, T_N\}. \tag{1.1}$$

Under very simple conditions, namely that all components have the same distribution, $F(t)$, of failure time, and all failure times are independent, simple conditional probability arguments yield

$$P\{X_N > t\} = E_N[(1-F(t))^N] \tag{1.2}$$

$$\equiv g_N(\bar{F}(t))$$

where g_N is the generating function of the number of faults originally sown in the system, and $\bar{F}(t) = 1-F(t)$ is the survival time distribution, per fault. It is easy to see that $P\{X_N = \infty\} = P\{N = 0\}$, possibly > 0, so the derived distribution of X_N is quite possibly dishonest. Note that while in general explicit expressions for expectations cannot be obtained (may not even exist), such summaries as the median, 90% point, etc., may if $g_N(\bar{F}(t))$ can be explicitly inverted, e.g. for N ~ Poisson and F ~ Exponential. The simplistic assumption of the model may be relaxed, allowing for different T_i distributions, dependence, and so on, and an additional random death time, D, applying to the total system can be introduced to induce eventual failure (of physical equipment), or biological death in finite time. There will be less analytical tractability,

but simulation may be used to assess system behavior. Statistical estimation problems may be addressed as well; a suitable version of (1.2) will provide a likelihood function.

Another example of the applicability of a simple conditioning argument is the following.

Situation: When an individual speaks on a telephone or telecommunication channel the conversation is an alternating sequence of talk-spurts and pauses. Similarly, a job being processed on a computer goes through an alternating sequence of CPU (compute) times and IO (input-output) times. Model the total time of the conversation or job processing time, and particularly the joint distribution of busy and idle segments.

Model 1: Let $\{X_i,\ i = 1,2,\ \ldots,\ K\}$ and $\{Y_i,\ i = 1,2,\ \ldots,\ K\}$ be the durations of talk spurts and pauses, respectively, and let K be the number of each. The simplest model assumes $\{X_i\}$ and $\{Y_i\}$ to be independently and identically distributed (IID) sequences of rv, and themselves to be conditionally independent, given K, also a rv. The joint distribution of total talking (or processing) time, X, and total pause time (IO time), Y, is thus, by simple conditioning,

$$P\{X \leq x,\ Y \leq y\} = \sum_{k=1}^{\infty} F_X(x)^{k*} \cdot F_Y(y)^{k*}\ P\{K=k\} \tag{1.3}$$

The joint Laplace transform is $(s_1,\ s_2,\ \geq 0)$

$$E\left[e^{-s_1 X}\ e^{-s_2 Y}\right] = \sum_{k=1}^{\infty} [\hat{F}_X(s_1)\hat{F}_Y(s_2)]^k\ P\{K=k\} \tag{1.4}$$

$$= g_K[\hat{F}_X(s_1)\hat{F}_Y(s_2)].$$

where g_K is the generating function of K. Put $s_1 = s_2 = s$ to recover the transform of $X+Y = L$, the total conversation length. In case

$X_i \sim \text{Expon}(\lambda)$ and $Y_i \sim \text{Expon}(\mu)$ and $K \sim \text{Geom}(\alpha)$, independent:

$$E\left[e^{-s_1 X} e^{-s_2 Y}\right] = \sum_{k=1}^{\infty} \left(\frac{\lambda}{\lambda+s_1}\right)^k \left(\frac{\mu}{\mu+s_2}\right)^k (1-\alpha)\alpha^{k-1} \tag{1.5}$$

$$= \frac{\lambda\mu(1-\alpha)}{(\lambda+s_1)(\mu+s_2) - \lambda\mu\alpha}$$

and

$$E[X] = [\lambda(1-\alpha)]^{-1}, \quad E[Y] = [\mu(1-\alpha)]^{-1} \tag{1.6}$$

Furthermore ($\bar{\alpha} = 1-\alpha$)!

$$E[L] = E[X] + E[Y] = \frac{\lambda+\mu}{\lambda\mu\bar{\alpha}} \tag{1.7}$$

and

$$\text{Var}[L] = \left(E[L]\right)^2 - \frac{2}{\lambda\mu\bar{\alpha}} \tag{1.8}$$

Notice also that the mechanism of <u>randomization of a sum</u>, or <u>mixing</u> (see Feller [1966]) which has given (1.3) may be used to generate families of bivariate (multivariate) exponential distributions for other modeling purposes.

<u>Model 2</u>: A plausible alternative to the above model assumes X_i and Y_i are not independent, being possibly positively correlated -- a long talkspurt tending to result in a long pause (response by conversationalist). Most simply,

$Y_i = \beta X_i$, $\beta > 0$. Then again transform in the $X \sim \text{Expon}(\lambda)$ case to get

$$E\left[e^{-s_1 X} e^{-s_2 Y}\right] = \frac{\lambda(1-\alpha)}{s_1 + \beta s_2 + \lambda(1-\alpha)}$$

or if $\beta = \lambda/\mu$ which preserves the marginals of Model 1,

$$= \frac{\lambda(1-\alpha)}{s_1 + \frac{\lambda}{\mu} s_2 + \lambda(1-\alpha)} = \frac{\lambda\mu\bar{\alpha}}{s_1\mu + s_2\lambda + \lambda\mu\bar{\alpha}} \quad . \qquad (1.9)$$

It is now immediate that the marginal distribution of $X \sim \text{Expon}(\lambda\bar{\alpha})$, $Y \sim \text{Expon}(\mu\bar{\alpha})$ and that now the df of the total time, L, is simply $\text{Expon}\left(\frac{\lambda\mu\bar{\alpha}}{\lambda+\mu}\right)$ -- a much simpler form than that occurring in Model 1 above, which involves a Bessel function. The variance of L in Model 2, being $(E[L])^2$, is also larger than that for Model 1, see (1.8), suggesting that the former model has a longer tail, hence predicting a greater proportion of extremely long conversations.

The above illustrates that the same situation can easily give rise to two -- or more -- different models, depending upon the manner in which stochastic assumptions are introduced. At best, the introduction should be guided by observed data; at least, sensitivity analyses using different assumptions can outline the range of specification uncertainty.

(2) Models Involving Repeated Trials

A great many situations may be initially modeled in terms of repeated independent trials, where this means that on each of a possibly countably infinite number of occasions a trial (or experiment, or observation) is performed, with outcome X_i (possibly a vector random variable (rv)) on the ith trial; $\{X_i, i = 1, 2, ...\}$ are IID rv. Bernoulli trials are a prime example: flip a biased coin indefinitely; let I_i be one if a Head (Success) results, and zero if a Tail (Failure) otherwise, and assume that probability of success on any trial is independent of all previous outcomes. Equally, X_i may be the winnings on a bet at occasion i, with X_i in dollars and either positive or negative. Or X_i may even represent the increase or decrease in a common stock price on the New York Stock exchange, according to some observers.

It is convenient, but somewhat more questionable, to adopt the repeated trial model for modeling real operational and physical phenomena, yet it is often done uncritically.

For instance the lifetimes (times between failure) of computing equipment are often modeled by IID rv, repair times likewise, queueing system inter-arrivals and service times as well, inventory demand sizes, sizes of deposits of resources (petroleum) as well, ... the list is very long, and observational support for these assumptions is usually conspicuously lacking. The Attraction of the repeated trials model is mainly its mathematical tractability, which leads to elegant and appealing results.

Often a brief data analysis in terms of marginal distributions of observed X_i seems to provide justification. The simple repeated trials model cannot well represent, say, systematic daily changes in job inter-arrival times, or numbers of jobs per hour, at a computer center, seasonal effects on such computer system demand, or the influence of other variables such as the introduction of a new class of computer users upon a measure of computer loading. Also, the model does not well describe the sequence of daily rainfalls in a region, nor many other environmental variables. Some examples follow in which repeated trial models seem initially plausible, but which doubtless can stand improvement.

Situation: A structure is to be designed to withstand (wind-generated, or seismic or other) shocks. Question: how long will it survive the environment stresses, given its initial strength Z?

Model: Let the structure have strength Z (suitable units). If the ith year's maximum shock is X_i, assume $\{X_i\}$ to be IID with df $F_X(x)$. Then the time T until structure failure exceeds t ($t = 1, 2, \ldots$) iff $\forall X_i < Z$, $i = 1,2, \ldots, t$, so

$$P\{T > t|Z\} = [F_X(Z)]^t, \tag{2.1}$$

the geometric distribution,

while if Z itself is regarded as random

$$P\{T > t\} = E_Z\, P\{T > t|Z\} = E_Z\{[F_X(Z)]^t\} \neq \left(E_Z\{[F_X(Z)]\}\right)^t; \tag{2.2}$$

the unconditional distribution is a convex (probability) combination of exponentials. It will resemble an exponential, but often has an extended right tail. It is obviously important that the condition on equipment stress, Z, be removed at the appropriate stage -- at the end. Removal of the condition before each "strength test" would be appropriate only if the structure were completely repaired or replaced with another having the unchanged distribution of Z before each trial.

Note that assuming yearly maximum environmental events to be IID is intuitively plausible, but some statistical evidence exists for truly long-range correlation in weather data that may call even this assumption into question.

The Bernoulli counting process is an important special case of repeated trials, on each of which either Success (probability p) or Failure (probability $q = 1-p$) occurs; $\{N(t), t = 1,2, \ldots\}$ is the number of Successes in $(1,t]$). Times (number of trials) τ between successive Successes are IID and geometrically distributed ($P\{\tau = k\} = q^{k-1}p$). The number $N(t)$ of Successes in fixed time is Binomial (p, t), and is in turn approximately Normal (tp, tpq) as $t \to \infty$. Of course Bernoulli trials describe the outcomes of many other repeated trial situations: for instance, the number of jobs submitted to a processing facility requiring more than x time units of processing may be modeled as a Bernoulli counting process with $p = 1 - F_X(x)$, F_X being the distribution of job processing time.

Generalizations of Bernoulli trial situations may be (a) to variations of success probability with trial number ("time", or "space"): $P\{X_i = 1\} = p_i$. Here $E[N(t)] = \sum_{i=1}^{t} p_i$ and $\mathrm{Var}[N(t)] = \sum_{i=1}^{t} p_i q_i \leq t\bar{p}(1-\bar{p})$, where $\bar{p} = \frac{1}{t}\sum_{i=1}^{t} p_i$, so variability is under-represented if trial-to-trial probabilities change, but

a Normal approximation to N(t) may still hold. A second generalization is to (b) independently randomize p in the Binomial distribution, say according to a Beta distribution, creating the Beta-Binomial distribution. This has found useful in reliability modeling and in Bayesian inference. A third and important elaboration, (c), is to develop a statistical regression model for success probability p based conveniently on the logistic function:

$$p_i = \frac{e^{\alpha+\beta u_i}}{1+e^{\alpha+\beta u_i}} ; \qquad (2.3)$$

here u_i (possibly a vector) represents the influence of other factors upon success probability. Given observations of the form (I_i, u_i), where $I_i = 1$ indicates success on trial i, one can estimate α and β (vector) by maximum likelihood; see Cox [1969]. Generalizations to multiple-category situations are possible, and computational methods for parameter estimation and model assessment have been devised; see Pregibon [1981].

The distribution of the number of counts N(t) in t trials of a Bernoulli trials process can be computed by making use of a forward equation. Let

$$P_j(t) = P\{N(t) = j \mid N(0) = 0\}.$$

Then

$$P_j(t) = P_j(t-1)\cdot(1-p) + P_{j-1}(t-1)\cdot p; (0 \leq j \leq t); \qquad (2.4)$$

on the basis of conditioning on events that have happened up to t-1. One can generalize to non-stationary success probabilities easily:

$$P_j(t) = P_j(t-1)\cdot(1-p_t) + P_{j-1}(t-1)p_t \quad . \tag{2.5}$$

Initial conditions may be

$$P_o(0) = 1, \quad P_j(0) = 0, \quad j = 1,2, \ldots \quad .$$

A further generalization allows success probability to depend upon the number of previous successes; then

$$P_j(t) = P_j(t-1)(1-p_{j,t}) + P_{j-1}(t-1)p_{j-1,t} \quad , \tag{2.6}$$

and the distribution $P_j(t)$ can easily be computed recursively given the success probabilities

$$p_{j,t} = P\{N(t) = j+1 \mid N(t) = j\} \quad . \tag{2.7}$$

This is a preview of ideas of Markov chains, to be treated later. These expressions are introduced to suggest early that the answers to interesting and comparatively complex problems can be directly computed numerically (in this case iteratively, starting from the initial conditions). Closed-form expressions such as the Binomial distribution are handy, and Normal approximations are even handier, but one need not modify the facts merely for the sake of convenience.

(3) Sums of Repeated Trial (IID) Random Variables; "Large Deviations"

Models for total demand for physical inventory or for facility (computer) time often naturally involve sums of varying components, modelled as rv.; thus total demand from n sources, or over n time periods, is

$$S_n = X_1 + X_2 + \ldots + X_n. \tag{3.1}$$

If X_i is the (dollar) profit in the ith year for some enterprise, then a financial measure of success is

$$S_n(r) = \sum_{i=1}^{n} X_i r^i \tag{3.2}$$

where r is a discount rate ($0 < r \leq 1$).

Situation. A computer center experiences varying monthly demands, X_i for the ith month. Here are answers to several simple questions involving sums of X_is.

The expected yearly (n-period) demand is

$$E[S_n] = \sum_{i=1}^{n} E[X_i], \tag{3.3}$$

the sum of the expected monthly demands, and also

$$Var[S_n] = \sum_{i=1}^{n} Var[X_i] \tag{3.4}$$

provided the X_is are uncorrelated. Importantly, as $n \to \infty$

$$F_{S_n'}(x) \equiv P\left\{\frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} \leq x\right\} \approx \int_{-\infty}^{x} e^{-\frac{1}{2}z^2} \frac{dz}{\sqrt{2\pi}} \equiv \Phi(x), \tag{3.5}$$

i.e. S_n becomes approximately Normally distributed no matter what the distributions of X_i, by the Central Limit Theorem, provided the X_i components are all of about the same size (certainly if they all come independently from the same parent distribution with finite mean and variance). For smallish n or distinctly non-Normal components the approximation is improved by an Edgeworth expansion (Feller [1966]), wherein for the equal component example

$$F_{S_n}(x) = \Phi(x) + \left(\frac{\mu_3}{\sigma^3}\right) \cdot \frac{1}{6\sqrt{n}} (x^2-1) \frac{d\Phi}{dx} + R_n \qquad (3.6)$$

where $R_n = 0\left(\frac{1}{\sqrt{n}}\right)$, and the components are assumed to have densities. Here $\mu_3 = E\left[(X - E[X])^3\right]$ and the term $\frac{\mu_3}{\sigma^3} = \gamma_1$ is the conventional dimensionless skewness measure for a distribution, being zero for symmetric distributions (Normal), and being +2 for the Exponential. Additional terms involving kurtosis (4th moments) improve the approximation, but it is possible that Edgeworth numerical values can be "infeasible": the approximation can actually decrease with x in certain ranges. Nevertheless the Edgeworth series has been usefully applied, even to unequal component situations, for estimating the loss of capacity of an electric utility; see Levy & Kahn [1981].

A useful alternative is the method of large deviations, Feller [1966], Daniels [1954], and others. The ingenious idea is to tilt (or sculpture) the df. components so as to make a Normal approximation more effective at predicting the probability that $S_n > x$ for large x. For equal components with

d.f. $F(x)$ look at the tilted probability measure (assumed to exist for $s > 0$, which sometimes restricts the theoretical applicability):

$$V(dx) = \frac{e^{sx}F(dx)}{e^{\psi(s)}} , \tag{3.7}$$

$\psi(s) = \ell n\, \hat{F}(s) = \ell n\, E[e^{sX}]$ being a cumulant generating function for F, or X. Manipulations show that

$$P\{S_n > z\} \equiv \int_z^\infty F^{n*}(dz) = e^{n\psi(s)} \int_z^\infty e^{-sx}\, V^{n*}(dx), \tag{3.8}$$

and the idea is to approximate V^{n*} by a Normal centered at z, a feat that can be accomplished by choice of s. It turns out that it is necessary to solve (sometimes numerically) for $s(z)$ the equation

$$z = n\psi'(s) \tag{3.9}$$

in order that the mean of the approximating Normal be at z; the variance is $n\psi''(s)$. Finally,

$$P\{S_n > z\} \sim \exp n\{\psi[s(z)] - s(z)\psi'[s(z)] + \tfrac{1}{2} s^2(z)\psi''[s(z)]\} \times \frac{1}{\sqrt{2\pi}} \int_{s(z)\sqrt{n\psi''[s(z)]}}^\infty e^{-\frac{1}{2}v^2}\, dv \quad . \tag{3.10}$$

The above technique can also be applied to a compound Poisson model (n is replaced conditionally by $N(t)$, the counting process of a Poisson process, and the condition then removed). Such models are frequently employed in inventory studies; apparently the large deviations approximation has not been applied in that area.

(4) Bernoulli Trials and Poisson Process: Rare Events

Bernoulli Trials are a special case of the Repeated Trials model, with events occurring ("Success") or not permitted to occur ("Failure") at specific integer time points, often equally spaced. In practice the fixed intervals between trials may be largely arbitrary, and it is attractive to think of events occurring at any (real-valued) time; from this comes the Poisson process. One approach to the P.P. properties is to consider a B.T. process to operate over time t with unit time steps, and then refine the time steps (e.g. let t = 1 day and starting with possible demands at 15-minute intervals, then down to 7.5 minutes, then to 3.75 etc.) to create a sequence of B.T. models. The limit of the sequence after ultimate refinement describes the P.P.

Specifically, let $T(k)$ be the generic time between successes in the (kth) B.T. model with time steps $1/2^k$ $(k = 0,1,2, \ldots)$. This means that the actual number of steps in time t for B.T. model k is $2^k t$; correspondingly, let the probability of success per step be $p/2^k$. By conditioning on the first step's outcome this means that

$$E[T(k)] = 1/2^k + (1 - p/2^k) \cdot E[T(k)] \qquad (4.1)$$

so $E[T(k)] = \frac{1}{p}$ for every model, as should be true. Furthermore, as $k \to \infty$ so time steps become arbitrarily small,

$$P\{T(k) > t\} = (1 - p/2^k)^{t \cdot 2^k} \to e^{-tp} \qquad (4.2)$$

inter-event times become exponentially distributed in the (P.P.)

limit. Furthermore the number of P.P. events ("Successes") in time t have the Poisson distribution.

The P.P. is usefully invoked for many modelling purposes.

Situation. Consider a sequence of days on which demands for computer service (time) are made, and focus on the occurrence patterns of runs (uninterrupted sequences) of high-demand days. Question: what is the distribution of times between successive runs, and what is the distribution of the number of such runs in a fixed time t? It will turn out that if either the run lengths are long, or if the probability of a high-demand day is small, that runs tend to occur as a Poisson process if the time scale is appropriate.

Model. Begin by modelling individual high demand day occurrences as successes in B.T. Let $\tau_1(k)$ represent the time until the 1st occurrence of a run of length k, and, measured from the end of such a run, let $\tau_2(k)$, $\tau_3(k)$, ... $\tau_i(k)$,.. be the time until the 2nd, 3rd, ... ith, such run is realized. By the B.T. assumption $\{\tau_i(k), i = 1,2, ...\}$ is an IID sequence of rv. Then we can represent $\tau(k)$ by conditioning on the events that may occur in the first k trials:

$$\tau(k) = \begin{cases} k & \text{with prob.} \quad p^k \\ 1 + \tau'(k) & \text{with prob.} \quad q \\ \ldots & \\ j + \tau'(k) & \text{with prob.} \quad p^{j-1}q \\ \ldots & \\ k + \tau'(k) & \text{with prob.} \quad p^{k-1}q \end{cases} \tag{4.3}$$

where $\tau'(k)$ is an independent replica of any $\tau(k)$: the idea is that the process starts over once a failure occurs to spoil a run. Alternatively,

$$\tau(k) = \begin{cases} k & \text{with prob. } p^k \\ R(k) + \tau'(k) & \text{with prob. } 1-p^k \end{cases} \tag{4.4}$$

where

$$P\{R(k) = j\} = \frac{qp^{j-1}}{1-p^k}, \qquad j = 1,2, \ldots, k, \tag{4.5}$$

a truncated geometric distribution. From these come the generating function of $\tau(k)$, and in principle its distribution:

$$E\left[z^{\tau(k)}\right] = \frac{z^k p^k}{1 - qz\left[\frac{1-(pz)^k}{1-pz}\right]} . \tag{4.6}$$

Differentiation gives the mean

$$E[\tau(k)] = k + \frac{1}{p^k}\left[\frac{1}{1-p} - \frac{kp^k}{1-p^k}\right] \approx \frac{1}{p^k(1-p)} \tag{4.7}$$

the approximation holding if either $p \to 0$ or $k \to \infty$. In either case the run is a rare event.

While explicit inversion of the expression for $E[z^{\tau(k)}]$ is possible by use of partial fractions, the result is quite complicated. On the other hand, look for the distribution of

$$\tau^*(k) = \tau(k)/E[\tau(k)] \tag{4.8}$$

when $E[\tau(k)]$ becomes large. The expectation

$$E[e^{-s\tau^*(k)}] = E\left[e^{-s\tau(k)/E[\tau(k)]}\right] \qquad (4.9)$$

can be obtained from the generating function (by putting $z = \exp[-sp^k q]$); next let either $p \to 0$ (rare individual events) or $k \to \infty$ (long runs) to find that this transform converges to $(1+s)^{-1}$. Then by the unicity theorem for transforms (Feller [1966]) the normalized rv $\tau^*(k)$ is approximately unit exponentially distributed, i.e.

$$P\{\tau(k) \leq t\, E[\tau(k)]\} \simeq 1 - e^{-t} \qquad (4.10)$$

and furthermore the distribution of the number of k-runs in time t, $N_k(t)$, is approximately Poisson. Deviation from the Poisson (indicated by over-variance) may signify that the underlying demand generating process is inhomogeneous or cluster-prone in time, and that extra facilities are required to reduce backlogs. Examination of runs is one way to check the validity of the basic modeling assumption of Bernoulli trials.

Similar limiting arguments simplify other situations involving rare events that are generated by even more complicated processes. See work on first-passage times for combinations of random loads by Gaver, Jacobs and Latonche [1981].

(5) Markov Models: General Comments

The basic theory of Markov chains and processes, both in discrete and continuous time, is well introduced in standard texts such as Feller [1966], Chung [1967], Karlin & Taylor [1975] Kleinrock[1976], and needs no systematic coverage, only review and illustration. By way of review, recollect the ideas of various possible state space definitions: integers, integer and real numbers ("ages"), real numbers (e.g. virtual waiting times in queues); times (index sets) either discrete and equally-spaced or imbedded or continuous time; Markov property defined by conditional probabilities ("The future is independent of the past, given the present"). Carry on to matrix representation of the state probabilities after t (0,1,2,...) time steps, forward and backward Chapman-Kolmogorov equations, generalize to discrete state Markov chain in continuous time with exponential sojourns in states, state classification emphasizing irreducible chains and transient chains (with at least one absorbing barrier), recurrent events and first-passage times and absorption probabilities, generating functions and other transforms.

Simple Markovian assumptions, i.e. that a scalar state rv $X(t)$, where t is time or space, is Markov, introduce dependence in a plausible and tractable manner. Usually it is necessary to assume, for example, that the one-step transition probabilities (discrete state, discrete time):

$$p_{ij} = P\{X(t) = j \mid X(t-1) = i\}, \tag{5.1}$$

are time-homogeneous in order to obtain explicit neat solutions. Analogous assumptions must be made about discrete-state Markov processes in continuous time, wherein λ_i is the rate of departure from state i (exponentially distributed sojourn time parameter), and p_{ij} is the corresponding probability of move from i to destination j. Of course a known deterministic time dependence, involving daily or weekly cycles, and trends can be dealt with by numerically multiplying the transition probability matrices.

More irregular changes in process behaviour can be represented as the effect of randomly changing external events, or random environments for short. In such models the actual primary process transition parameters (e.g. p_{ij}, or λ_i) change in time under the influence of such environmental factors as seismic vibration, temperature and humidity, ocean sea state, wind speed or other meteorological effects, or variations in personnel effectiveness and propensity for errors. Random environment models conveniently postulate that environmental changes induce simple discrete-state Markovian behavior on the basic or primary process parameters; of most interest are parameter changes that occur more slowly than do state changes in the basic process.

Markov modelling of real situations usually involves simplifications at certain crucial states. Even then, the answers to interesting questions may require extensive computing or simulation. Astute choices of sub-models or component models, e.g. the use of "phase-type" distributions for representing arrival and departure processes in queues can be of help, as

can the recognition (or plausible imposition) and exploitation of special structure; see Neuts [1981].

(6) Some Markov Process Problems and Models

Here are some illustrative situations and corresponding Markov chain models.

Situation (Queueing in discrete time). A servicing facility, e.g. a computer system or a programming (or other) consultant, or a communication channel, experiences single customer arrivals in a random fashion; arrivals enter at the discrete times 0,1,2,3, ... only, and service completions occur only at such times. Discuss the nature of the delays and backlogs that occur.

Model 1. Let the probability of a single arrival at time (epoch) t be $a_i(t)$, where i refers to the number present at that epoch. Each arrival must wait at least one time period before discharge, even if it immediately enters service upon arrival. Let $d_i(t)$ be the probability that an arrival that has been in service at t actually departs at $t+1$. Now let $X(t)$ denote the number of arrivals in the system who have not yet completed service at time t. Model $\{X(t)\}$ as a Markov chain with the following one-step transition probabilities:

$$\begin{aligned} p_{i,i+1}(t) &= P\{X(t+1) = i+1 | X(t) = i\} = [1-d_i(t)]a_i(t) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , \ i \geq 1 \\ p_{i,i-1}(t) &= P\{X(t+1) = i-1 | X(t) = i\} = d_i(t)[1-a_i(t)] \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (6.1) \\ p_{0,1}(t) &= a_0(t) \\ p_{0,0}(t) &= 1-a_0(t) \end{aligned}$$

$$p_{ii}(t) = 1 - \{[1-d_i(t)]a_i(t) + d_i(t)[1-a_i(t)]\}$$

$$p_{ij}(t) = 0 \quad \text{otherwise;}$$

If the number in the system is $\leq I$, so the state space is finite, and $p_{I,I+1}(t) = 0$ and $p_{I,I-1}(t) = d_I(t)$ then the probability distribution of $X(t)$ for any t can be obtained by numerically multiplying the one-step transition matrices, $\underline{p}(t)$, with elements given above:

$$P_{ij}(t) = P\{X(t) = j \mid X(0) = i\} = \text{element}$$

in ith row, jth column of

$$\underline{P}(t) = \prod_{t'=0}^{t} \underline{p}(t') \tag{6.2}$$

This can be done especially easily in APL if the process is time-homogeneous, i.e. $a_i(t) = a_i$, $d_i(t) = d_i$ independent of elapsed time. Explicit analytical solutions can rarely be found for non-time-homogeneous cases, let alone for time homogeneous cases. If they were available, the solutions would generally be very complicated and difficult to interpret.

Model 1-A. Specialize the above to let $a_i(t) = a > 0$ and $d_i(t) = d > 0$. If the maximum number in the system is I, there is a stationary solution; put $s = \frac{a\bar{d}}{\bar{a}d}$, $\bar{a} = 1-a$, $\bar{d} = 1-d$:

$$\begin{aligned}
\pi_0 &= \frac{(d-a)\bar{d}}{d\bar{d}-a\bar{a}s^I} \\
&\cdots \\
\pi_j &= \frac{(d-a)s^j}{d\bar{d}-a\bar{a}s^I} \\
&\cdots \\
\pi_I &= \frac{(d-a)\bar{a}s^I}{d\bar{d}-a\bar{a}s^I}
\end{aligned} \tag{6.3}$$

Furthermore, if $s<1$ then the process tends to drift towards 0 and even if there is no upper bound on system states the process is irreducible and ergodic so the long-run distribution is

$$\pi_j = \frac{\bar{d}-a}{d\bar{d}} s^j, \qquad j = 1,2, \ldots \tag{6.4}$$

a modified geometric distribution, some form of which so often appears in queueing problems. The above simple special case is well-known, but can be useful for checking the accuracy of computer programs used to compute numerical solutions to the time-dependent case.

Model 2. Let $a_{i,b}(t)$ be the probability that at time t there occurs a bunch of arrivals of size $b (b = 0,1,2, \ldots)$, given that i are awaiting service and will not make further demands. For example, suppose there are I total customers, e.g. computer terminals accessing a central facility, and that each applies for service independently with probability $\alpha(t)$, provided that it is not undergoing service. Then

$$a_{i,b}(t) = \binom{I-1}{b}[\alpha(t)]^b[1-\alpha(t)]^{I-i-b},$$

and for $k \geq 0$,

$$p_{i,i+k}(t) = \bar{d}_i(t)a_{i,k}(t) + d_i(t)a_{i,k+1}(t), \tag{6.5}$$

while

$$p_{i,i-1}(t) = a_{i,0}(t)d_i(t).$$

Again the probability distribution of $X(t)$ can be numerically computed.

Model 2-A. Suppose the arrivals are caused by a common event (a "common-cause" in engineering parlance). This might be the occurrence of an earthquake of large magnitude, or other environmental shock. Let such an event occur at time t with probability $c(t)$; let the probability that b arrivals (demands for service) occur as a result be conditionally binomially distributed with parameter θ. Then

$$a_{i,b}(t) = c(t).\binom{I-i}{b}\theta^{b}[1-\theta]^{I-i-b} \qquad (6.6)$$

This can again be used to form one-step transition probabilities, and to calculate state probabilities at any time. The present model allows for a catastrophic shock situation: if $\theta = 1$ then all outstanding customers simultaneously demand service, i.e. $I-i$ arrivals occur simultaneously. This differs from Model 2.

Situation (Queueing with breakdown of service or preemptive priorities). Suppose a single server, e.g. computer facility, or data transmission channel, is confronted by a random arrival stream of basic service demands. These demands may be characterized by their service times, or work request durations such as the times required to transmit single bodies of data or digitized messages. In addition, these services may be effectively prolonged by the occurrence of interruptions, e.g. from internal server breakdowns resulting in temporary processor unavailability, or

from environmental noise or even intentional jamming. How is the queue size and waiting time of demands affected by such interruptions? What steps can be taken to reduce the interruption effects?

Model. Nearly all classical queueing theory is most conveniently developed if the service times of the individual demands are independently and identically distributed (IID); see however Jacobs [1978, 1980] for discussion of a model involving correlation effects. If service times are to be interrupted and repeated, or alternatively resumed, an interruption process that preserves the IID character of the basic service times allows nearly direct adaptation of conventional theory; such a process is one that requires exponentially distributed ("memoryless") periods between successive interruptions, and this will be assumed. Checks of the sensitivity of results to this reasonable assumption can be made by simulation.

If interruptions of IID duration X (df $F_X(x)$) occur at IID Expon(λ_H) intervals, then the time to complete the ith basic service (low-priority) is, provided service can resume after each service

$$C_i = S_i + X_1 + X_2 + \cdots + X_{I(S_i)} , \qquad (6.7)$$

where S_i is the ith basic or low priority service time (df $F_s(x)$), X_j is the duration of the jth interruption, and $I(S_i)$ is the number of interruptions that occur during S_i.

Given S_i, $I(S_i)$ is Poisson (λ_H), and the Laplace-Stieltjes transform of C_i is, in terms of

$$\hat{F}_C(s) = \hat{F}_S[s+\lambda_H\{1-\hat{F}_X(s)\}] \tag{6.8}$$

and hence

$$\begin{aligned} E[C] &= E[S]\{1+\lambda_H E[X]\} \\ E[C^2] &= E[S^2]\{1+\lambda_H E[X]\}^2 + E[S]\lambda_H E[X^2]. \end{aligned} \tag{6.9}$$

If, on the other hand, basic services that are interrupted must begin again from scratch, i.e. services must <u>repeat</u>, then the <u>ith</u> completion time becomes

$$C_i = S_i + X_1 + S_1' + X_2 + S_2' + \ldots + X_{I(S_i)} + S_{I(S_i)}' \tag{6.10}$$

where S_j' is the <u>jth</u> interrupted basic service time that must be repeated. The L.-S. transform is

$$\hat{F}_C(s) = E\left\{\frac{e^{-(\lambda_H+s)S}}{1-E[e^{-sX}]\,\frac{\lambda_H}{\lambda_H+s}\left[1-e^{-(s+\lambda_H)S}\right]}\right\}, \tag{6.11}$$

and by differentiation of the latter expression,

$$E[C] = \left\{E\left[e^{\lambda_H S}\right]-1\right\}\left\{E[X] + \frac{1}{\lambda_H}\right\} \tag{6.12}$$

$$E[C^2] = 2E\left[\left(e^{\lambda_H S}-1\right)^2\right]\left(E[X] + \frac{1}{\lambda_H}\right)^2 + 2E\left[Se^{\lambda_H S}\right]\left(E[X] + \frac{1}{\lambda_H}\right)$$

$$+ \left(E\left(e^{\lambda_H S}\right)-1\right)\left(E[X^2] + 2E[X]\cdot\frac{1}{\lambda_H} + \frac{2}{\lambda_H^2}\right)$$

In order to assess queueing delay, look at the process, $\{N_d, d=0,1,2,\ldots$ describing the number of basic demands at the server at the instants just following departures; $\{N_d,\ d = 1,2,\ \ldots\}$ is an embedded Markov chain provided basic arrivals are (compound) Poisson. The

$$\begin{aligned} N_{d+1} &= H_d + A(C_{d+1}) \qquad \text{if } N_d = 0 \\ &= N_d + A(C_{d+1}) - 1 \qquad \text{if } N_d \geq 1 \ . \end{aligned} \tag{6.13}$$

Here H_d is the number of basic (low-priority) demands made at the beginning of a basic service busy period initiated by the appearance of a high-priority demand, and $A(C_d)$ is the number of basic demands made during the dth basic completion time. Express an arbitrary H as follows

$$H = \begin{cases} 0 & \text{with probability } \dfrac{\lambda_L}{\lambda_L+\lambda_H} \\ A(X) & \text{with probability } \dfrac{\lambda_H}{\lambda_L+\lambda_H} \end{cases} . \tag{6.14}$$

It follows by conditional expectations that the embedded chain is ergodic if $E[A(C)] < 1$, and that then the long-run probability of system emptiness at an embedded time point is

$$p_o \equiv \lim_{d\to\infty} P\{N_d=0\} = \frac{1 - E[A(C)]}{E[H]} \tag{6.15}$$

and the long-run expected occupancy is

$$E[N] = \frac{1}{2(1-E[A(C)])} \left\{ E[(H + A(C))^2]p_0 + E[(A(C)-1)^2](1-p_0) \right\}. \tag{6.16}$$

Delay can then be estimated by use of Little's formula. A version of the above formulas correct in continuous time may be found using

results of Gaver [1962]; the difference between embedded and continuous time becomes comparatively negligible if the basic traffic intensity E[A(C)] is close to, but below unity ("heavy traffic"). Since computer system monitoring devices sample system state at the moment an event occurs (e.g. at a departure instant) a theoretical account of the queue at such moments (imbedded times) is of direct interest.

An alternative approach to the long-run distribution of delay of an arriving basic demand is by way of Wald's identity or martingales; see Feller [1966]. This will actually handle waiting-times when the basic service inter-arrival intervals are IID, but otherwise arbitrarily distributed. Still another approach is via the Takacs-type integro-differential equation; see Kleinrock [1976] for an account.

The previous situations have been discussed in terms of long-run probabilities. Frequently questions involving the time until system failure (or restitution to operational condition) are more important.

Situation (Redundant repairable systems). A particular system function can be performed if at least k out of n system components ("machines") function. For example, electric power is available if at least one generator is working out of two that are installed. Suppose that the system components are all operative initially but fail randomly; failures are immediately detected, but repairs are of random duration, so several machines can be down, all awaiting repair completion. Question: how long will it be until $\ell = n-k+1$ components are simultaneously in a failed condition? The time until this occurrence is the time to failure of a k-out-of-n system.

Model 1. It is now convenient (but possibly unrealistic!) to assume that the machines fail independently and after exponentially distributed times in operation, each with rate λ. This simplification may be relaxed, but at the price of expanding the state space. Assume too that the repairs occur in a Markovian manner, e.g. (but not necessarily) at rate $\mu \cdot \text{Min}(N(t),R(t))$, where $N(t)$ is the number of machines failed and down for repair at t, and $R(t)$ is the number of repairmen on duty. This is the classical machine repairman problem; see Feller [1966], and Cox and Smith [1962]. Usually $R(t)=r$, a constant, although provisions may be made for automatically increasing repair effort when redundancy reserves become dangerously low. In other words, $N(t)$ is a simple birth-and-death Markov process, wherein jumps in state, $N(t)$, occur at exponentially distributed intervals or sojourn times, S_i for the sojourn in state i,

transitioning always to neighboring states. In the present situation as $\Delta \to 0$

$$\begin{aligned} P\{N(t+\Delta) = i+1 \mid N(t) = i\} &= \lambda(n-i)\Delta + o(\Delta) \\ &= \lambda_i \Delta + o(\Delta) \\ P\{N(t+\Delta) = i-1 \mid N(t) = i\} &= \mu \min(i,r)\Delta + o(\Delta) \\ &= \mu_i \Delta + o(\Delta), \end{aligned} \tag{6.17}$$

abbreviating the general transition rates to λ_i and μ_i.

In order for system state to reach $i+1$ from i for the first time it must either do so on the first transition out of state i, or else drop back to i-1, return to i and try again. Thus if U_i is the local first passage time from i to $i+1$:

$$U_i = \inf\{t: \ N(t) = i+1 \mid N(t) = i\}. \tag{6.18}$$

write

$$U_i = S_i + \begin{cases} 0 & \text{with probability } p_{i,i+1} = \dfrac{\lambda_i}{\lambda_i+\mu_i} \\ U'_{i-1} + U'_i & \text{with probability } p_{i,i-1} = \dfrac{\mu_i}{\lambda_i+\mu_i} \end{cases} \tag{6.19}$$

where U'_i has the same distribution as U_i. The above representation allows immediate derivation of the Laplace-Stieltjes transform of U_i by conditional expectations. The result is

$$\begin{aligned} E\left[e^{-sU_i}\right] \equiv \psi_i(s) &= \frac{\lambda_i}{s+\lambda_i+\mu_i[1-\psi_{i-1}(s)]}, \quad i = 1,2, \ldots, \\ \psi_o(s) &= \frac{\lambda_o}{\lambda_o+s} \ . \end{aligned} \tag{6.20}$$

Furthermore, since the first-passage time from $N(t) = i$ to $j > i$ can, on the basis of Markovian assumptions, be expressed as

$$T_{ij} = U_i + U_{i+1} + \cdots + U_{j-1} \tag{6.21}$$

where $\{U_{i+k}, k = 0,1, \ldots\}$ are independent rv, the L.-S. transform of T_{ij} is

$$E\left[e^{-sT_{ij}}\right] = \prod_{k=i}^{j-1} \psi_k(s) \tag{6.22}$$

and the cumulants (moment-like quantities; see Cramer [1946]) can be expressed in terms of those of the U_i. Here are a few moments of the U_i; recursively expressed and hence easily computed:

$$E[U_i] = \frac{1}{\lambda_i}\left\{1 + \mu_i E[U_{i-1}]\right\}.$$

$$E[U_i^2] = \frac{2}{\lambda_i^2}\left\{1 + \mu_i E[U_{i-1}]\right\}^2 + \frac{\mu_i}{\lambda_i} E[U_{i-1}^2]$$

$$E[U_i^3] = \frac{6}{\lambda_i^3}\left\{1 + \mu_i E[U_{i-1}]\right\}^3 + \frac{6\mu_i}{\lambda_i^2}\left\{1 + \mu_i E[U_{i-1}]\right\}E[U_{i-1}^2] + \frac{\mu_i}{\lambda_i} E[U_i^3] \tag{6.23}$$

and

$$E[U_i^4] = \frac{24}{\lambda_i^4}\left\{1 + \mu_i E[U_{i-1}]\right\}^4 + \frac{36\mu_i}{\lambda_i^3}\left\{1 + \mu_i E[U_{i-1}]\right\}^2 E[U_{i-1}^2] + \frac{6\mu_i^2}{\lambda_i^2}\left(E[U_{i-1}^2]\right)^2 + \frac{8\mu_i}{\lambda_i^2}\left\{1 + \mu_i E[U_{i-1}]\right\}E[U_{i-1}^3] + \frac{\mu_i}{\lambda_i} E[U_{i-1}^4].$$

From these, standard variance, skewness, and kurtosis measures can be easily computed. For the repairman model discussed initially it can be shown that if the expected time to system failure, $E[T_{o\ell}]$, is "long" then $T^{\#}_{o\ell}/E[T_{o\ell}]$ resembles an Expon(1) rv.

Model 2. (Catastrophic failures). Suppose that in addition to the independent random failures there is a catastrophic event that "kills" all operative machines simultaneously; let it occur after a time $C \sim \text{Expon}(\nu)$. Then the system failure time $T^{*}_{o\ell}$ has distribution given by

$$P\left\{T^{*}_{o\ell} > t\right\} = P\left\{T_{o\ell} > t\right\} e^{-\nu t} \tag{6.24}$$

and from this

$$E[e^{-sT^{*}_{o\ell}}] = \frac{1-E\left[e^{-(s+\nu)T_{o\ell}}\right]}{s+\nu} \tag{6.25}$$

from which moments can be generated; see Chu and Gaver [1977]. It is sobering to note that if ν^{-1}, the mean time to catastrophe occurrence, becomes small or even comparable to $E[T_{o\ell}]$, then the mean time to redundant system failure is essentially ν^{-1}, and redundancy alone may not improve system reliability.

Model 3 (Simultaneous repair). If the system is not under constant surveillance, but instead is inspected at random times (rate μ) and then repaired in negligible time, the number of down machines at t may jump essentially instantaneously, either to zero (perfect and rapid repair), or to some lower point (imperfect repair). In this case the basic "skip-free up"

character is retained, but now

$$U_i = S_i + \begin{cases} 0 & \text{with probability } p_{i,i+1} \\ U_i' + U_{i-1}' + U_{i-2}' + \ldots + U_{i-j}' & \text{with probability } p_{i,i-j}, \\ & j = 0,1,2, \ldots i. \end{cases} \tag{6.26}$$

Note that if $j = 0$ then repair is completely ineffective

Conditional expectations now give the L.-S. transform

$$E[e^{-sU_i}] \equiv \psi_i(s) = \frac{\eta_i(s)p_{i,i+1}}{1-\eta_i(s)\sum_{j=1}^{i}\left(\prod_{r=1}^{j}\psi_{i-r}\right)p_{i,i-j}}, \tag{6.27}$$

where here $\eta_i(s) = E[e^{-sS_i}] = \alpha_i(\alpha_i+s)^{-1}$. To specialize this to a repair model, introduce a Binomial distribution for successful repairs:

$$\begin{aligned} \alpha_i &= \lambda(n-i) + \mu \\ p_{i,i+1} &= \lambda(n-i)\,\alpha_i^{-1} \\ p_{i,i-j} &= \mu\alpha_i^{-1}\binom{i}{j}\rho^j(1-\rho)^{i-j}, \qquad j = 0,1,2, \ldots, i. \end{aligned} \tag{6.28}$$

where ρ represents the probability that an individual down machine is indeed repaired just after inspection (neglect the duration of repair times). The Binomial model assumes that repair success is independent across machines, which may be inappropriate in case similar causes give rise to the failures. Differentiation or direct expectations yield moments of U_i and eventually of $T_{o\ell}$.

(7) Diffusion and Fluid Approximation

While classical discrete state space Markov process ideas can often be used to model some quite interesting situations, the analytical results obtained frequently emerge only in terms of incomprehensible transforms, or in other somewhat obscure form. Not infrequently the difficulty that induces complexity can be traced back to the influence of boundaries upon the process transitions. If one examines a rather heavily loaded or congested service system, however, it is apparent, first, that the state changes may appear almost negligibly small relative to the system state magnitude (e.g. length of queue) itself, and, second, that annoying boundaries, particularly that at zero, are visited infrequently --- although their influence may still be crucial. These remarks hold true not only for simple one-dimensional processes, such as those used to describe congestion at a single servicing facility, but also for much more complex situations involving the interaction of several servicing processes.

An attractive approach to problems involving many customer arrivals occurring rapidly and generating considerable queueing is, then, to treat them by the method of diffusion process approximation. For details concerning the rigorous details of diffusion mathematics see Feller [1966]; in brief summary recall that a diffusion is a possibly vector-valued Markov process on the real numbers that typically moves continuously, governed by a drift (infinitesimal mean) and a diffusion (infinitesimal variance) parameter.

This approximation has been employed by Kobayashi [ ] for certain cyclic networks of queues such as are encountered in multiprogramming computer systems. See also Gaver and Shedler [1973], and the important work of M. Reiman [1982], and particularly G. Newell [1979] and also McNeil and Schach [ ]. In this section the use of diffusion will be briefly illustrated, and some experience with the results will be recounted.

Situation (Waiting time or backlog at one server). Suppose a single servicing facility is confronted by random arrivals that bring with them contributions to work load, expressed as required processing times. If the facility processes them in order of arrival, what is the backlog at time t? The facility is heavily loaded, so that it is seldom idle. It never turns away customers, i.e. infinite buffering is possible, nor do long delays discourage those waiting, causing defections or balking.

Model. Assume that arrivals occur in a Poisson (λ) process, and that the generic processing time S has df $G_S(y)$; successive processing times are IID. This model has been studied by Takacs [ ] who derived an integro-differential equation for the df of backlog or virtual waiting time W(t). Although the formal solution of that equation can be obtained, it is in a somewhat complicated form, not conducive to immediate insights. It is tempting to take an alternative, somewhat heuristic approach. Intuitively, if $\rho = \lambda E[S]$(= expected total load increment per unit time) > 1 (= processing or output rate per unit time), the backlog grows at rate $\rho - 1 > 0$. Furthermore, the backlog process,

$W(t)$, "ignores the boundary" at $W = 0$ after a time, and eventually $W(t)$ appears approximately Normal/Gaussian over an interval $(t, t+\Delta)$, with mean $(\rho-1)\Delta$, and variance $\lambda E[S^2]\Delta$. Importantly, also, the process seems to grow by accumulating independent, nearly Gaussian, increments.

If $\rho<1$ some difficulties occur because $W=0$ is an impermeable reflecting boundary, but the basic scenario is still the same: if ρ is close to unity $W(t)$ moves in nearly Gaussian increments, but occasionally interacts with the boundary at zero. Question: what is the long-run behavior of the delay in such a process?

The Takacs (or forward Kolmogorov) equation for the "exact" process is

$$-\frac{\partial F}{\partial x} + \frac{\partial F}{\partial t} = -\lambda F + \lambda \int_0^x F(x-y,t)G_S(dy) \qquad (7.1)$$

where $F(x,t)$ is the df of $W(t)$; initial and boundary conditions are necessary but are suppressed.

If $F(x,t)$ is only appreciable when x is large, and if the magnitude of a typical S is also small compared to x then it becomes plausible to Taylor-expand the $F(x-y,t)$-term to three terms and integrate; the result is

$$\frac{\partial \tilde{F}}{\partial t} = (1-\lambda E[S]) \frac{\partial \tilde{F}}{\partial x} + \frac{\lambda E[S^2]}{2} \frac{\partial^2 \tilde{F}}{\partial x^2} \qquad (7.2)$$

which is the well-known forward partial differential equation for Brownian motion with drift. Impose the boundary condition that $\tilde{F}(x,t) = 0$ for $x < 0$, and the equation for $\tilde{F}$, an approximation

to F, the "true" distribution of W(t), emerges; the latter equation can be explicitly solved in terms of error functions and exponentials (see Newell [1971]) for all t; i.e. the transient solution is actually readily available. With some further effort one can impose an upper boundary at $\bar{x} > 0$ to represent a finite buffer size. The approximate steady-state ($t \to \infty$) distribution turns out to be

$$\tilde{F}(x) = \exp\left[\frac{-2(1-\rho)}{\lambda E[S^2]}\, x\right], \qquad \rho < 1 \tag{7.3}$$

which often is, for large ρ, usefully close to the behavior of F(x), the long-run solution to the Takacs equation.

Note that the parameters of the above differential equation can be obtained by equating infinitesimal mean and variance of the assumed Poisson arrival process to the corresponding quantities for the diffusion. It is interesting that, when available, a martingale approach to the problem, of Gaver and Shedler [1973], produces a different exponent that yields better approximations for moderate traffic intensities, especially if the service time distribution is very long tailed (more skewed than the exponential).

Turn now to a more complex example, involving the interaction of two traffic dreams.

Situation. At a node of a communication network there are a total of c+v channels (servers); voice messages are exclusively assigned to the v channels, and data messages are assigned to the c channels, but may also utilize any unused voice

channel capacity under the stipulation that voice has preemptive priority and may displace any data encroaching on its (v-channel) territory. Question: what is the nature of the delay experienced by the data, which is allowed to queue up in a buffer indefinitely?

Model (Markov assumptions). Voice traffic is Poisson (λ) with Expon (μ) service times; voice is a loss system, so immediately the steady-state voice loss rate can be calculated using the Erlang-B formula. The data, however, operates in a random service environment modified by voice needs. Data arrives in an independent Poisson (δ) process, with independent Expon (η) service times. Typically $\delta >> \lambda$, and $\eta >> \mu$. Data, with state variable $X(t)$, is an M/M/S system where $S = c+v - V(t)$, $V(t)$ being the number of voice messages in service. Clearly $\{X(t), V(t)\}$ is a bivariate Markov process, but one difficult to analyze exactly; see references in Lehoczky and Gaver [1981] for other approaches to the analysis.

Now typically η/μ (= Data arrival rate per voice service time) is very large, possibly 10^4. Furthermore often $\rho_d \equiv \delta/\eta > c$, so some voice channel usage by data is necessary in order that all data be handled and there is not an evergrowing queue. The appropriate traffic intensity parameter for the system is seen to be

$$\rho = [\rho_d + \rho_v (1-q)](c+v)^{-1} \qquad (7.4)$$

where $\rho_v = \lambda/\mu$, and $q = \dfrac{\rho_v^v/v!}{\sum_{j=0}^{v} \rho_v^j/j!}$,

is the probability that the voice system rejects an arriving voice message. If ρ becomes large under such circumstances, Gaver and Lehoczky [1982] show that $X(t)$ behaves like a Wiener process with reflecting boundary, precisely as was mentioned in the previous example. Actually Lehoczky and Gaver [1982] assumes that data input acts as a fluid, with no variability.

By further, more intricate, methods involving the convergence of semigroups of operators developed by Burman [1979], it is shown in Lehoczky and Gaver [1981] that the long-run distribution of $X(t)$ is exponential with mean

$$\rho_d + \frac{\eta}{\mu\rho_r} \frac{\sum_{i=0}^{v-1} (T_i^2/\pi_i)}{(c+v)(1-\rho)} \qquad (7.5)$$

where

$$\pi_i = \rho_v^i/(i!)/\left[\sum_{i=0}^{v} \rho_v^i/(i!)\right]$$

and

$$T_k = \sum_{i=0}^{k} \pi_i (i-\rho_v(1-q)) \ .$$

Numerical work indicates that the diffusion approximation is reasonably accurate if the traffic intensity is quite high, say if $\rho \geq 0.95$; otherwise, for smaller s, the accuracy is not as high. It would not be surprising if a refined method for fitting drift and diffusion coefficients would lead to improved results. Finally, the difference between the refined treatment of basic data as a Poisson, and the simpler treatment by a fluid (yielding a model that can be solved exactly) resides

merely in the addition of the term ρ_d in the numerator of the expression for the mean.

The diffusion approximation can be utilized to evaluate another interesting measure of system effectiveness, namely the expected total waiting time, in job or data-packet hours, expended during a busy period for data traffic that starts with x present.

Model. Let $A(x)$ be the expected total waiting time during a busy period when the initial number of jobs is $X(0) = x > 0$. Condition on process change, $Z(\Delta)$, during the initial short time period $(0,\Delta)$ to get

$$E\left\{\int_0^\infty X(t')dt' \,\middle|\, X(0) = x, \quad X(\Delta) = x + Z(\Delta)\right\} \equiv$$

$$A\big(x;Z(\Delta)\big) = x\Delta + A\big(x+Z(\Delta)\big) + o(\Delta) \ . \tag{7.6}$$

Now Taylor-expand and remove the condition on $Z(\Delta)$:

$$A(x) = x\Delta + A(x) + A_x(x)\mu(x)\Delta + A_{xx}(x) \cdot \frac{\sigma^2(x)}{2}\Delta + o(\Delta) \tag{7.7}$$

or, collecting terms in Δ and letting $\Delta \to 0$,

$$0 = x + \mu(x)A_x(x) + \frac{1}{2}\sigma^2_{(x)} A_{xx}(x), \tag{7.8}$$

to be solved subject to $A(0) = 0$; clearly restrictions on $\mu(x)$ are necessary in order that $A(x)$ be finite.

If $\mu(x) = \mu < 0$ $\sigma^2(x) = \sigma^2$, both independent of x, the equation can be solved directly to give

$$A(x) = \frac{1}{-2\mu} x^2 + \frac{\sigma^2}{2\mu^2} x \quad ; \tag{7.9}$$

$\mu<0$ when traffic intensity $\rho<1$. A similar expression for compound Poisson(λ) inputs is

$$A(x) = \frac{x^2}{2(1-\rho)} + \frac{x\, E[A^2]}{2(1-\rho)^2} , \qquad \rho<1, \tag{7.10}$$

where A represents the number of items (packets) arising at a single request; $\rho = \lambda E[A]$. The moral is that variability (measured by σ^2 or $E[A^2]$) can greatly increase expected total waiting time, particularly when system loading is high (ρ close to unity).

For application of this "area under a random path" to discussing total wait during a road traffic jam see Gaver [1969]. See also McNeil [1970] for generalizations. The same backward argument is well-adapted also to studying problems of optimum investment decisions.

8. Renewal-Theoretic Modeling

Ideas of renewal theory and recurrent events are extremely useful for many purposes in stochastic modeling. Recognition of the occurrence of one or more recurrent events or "renewal points" in the development of a model process points the way to writing down simple forward or backward type equations for probabilities or expectations. Frequently analytical information can be extracted from such equations, particularly that relevant to long-time or other asymptotic process behavior. If more information is desired it can be obtained by use of transform techniques, by numerical computation, or by Monte Carlo simulation.

Mathematical definitions and properties of renewal processes are well presented by Feller [1966], Karlin and Taylor [1975], and Cox [1962], among others. There follow a few situations and suggested models based on renewal theory that illustrate the basic notions. We also comment on the relevance of the models and results obtained to real situations.

Situation. A machine, e.g. computer system or component thereof, or human operator, etc., operates properly for a period of time, fails, is restored (or restores itself) to service and operates properly again for a different time, fails again, and so on. Questions: how many failures are likely to occur in a given fixed period of time, say a year? The answer to such a question will help to guide decisions concerning logistics (necessary spare parts) and employment of repair personnel. How long a time will elapse until the k^{th} failure? Suppose the times to restore failures vary; what is the likelihood that a "chance"

user of the system will find the machine down for repair when he or she needs it, and how long will the wait to service restoration last? These are only a few of the many questions that might be asked.

Model. The classical renewal theory model for the situation described portulates that times between successive failures, $\{X_i, i = 1,2, \ldots\}$ are IID positive rv with distribution $F_X(x)$. That is X_1 is the time until the first event (here failure), and X_i is the elapsed time between the i-1st and ith event. X_i can be either a discrete random variable, so failures occur at regular intervals, say hourly, a continuous random variable, or a mixture. For the moment assume repairs to take a negligible time.

Mathematical results for this model are simplest and nicest when the IID assumption is fully exploited. Under that assumption (and even more generally) the counting process, N(t), giving the number of renewal events (failures in time t) has probability distribution

$$P\{N(t) = n\} = F_X^{n*}(t) - F_X^{(n+1)*}(t) \tag{8.1}$$

where * refers to convolution. For long time ($t \to \infty$) and under suitable mathematical restrictions

$$M(t) \overset{\sim}{=} E[N(t)] \overset{\sim}{-} \frac{t}{E[X]} \tag{8.2}$$

$$Var[N(t)] \overset{\sim}{-} \frac{Var[X]}{(E[X])^3}\, t$$

and $N(t) \sim$ Normal, with the above parameters. Of course exact analytical solutions can be obtained if sympathetic distributional models are assumed: taking $X \sim$ Expon. yields the Poisson distribution, and $X \sim$ Gamma or Erlang also produces rather neat closed-form solutions. Any discrete-time distribution for X can be numerically convolved, conveniently using APL. This helps to answer questions about the number of events in $(0, t)$, provided the IID assumption is palatable. If finite-time results are needed resort can be made to numerical summation, using a discrete time model, to approximation by a standard, tractable, distribution such as the Gamma followed by transform inversion, or by simulation.

In what follows we illustrate the diverse utility of backward conditioning arguments, leading to renewal integral equations.

Situation. (Incorrect repair possibly due to human error). Each time a repair is made there is the chance that it will be incorrect, and that the subsequent time to failure will be short. Suppose that incorrect repairs tend to bunch together in runs; describe the number of failures that occur over a (long) time period.

Notice that a similar situation describes a clustering scheme of arrivals to a repair facility or a communications center.

Model. Assume that the generic time to failure of the system is X when repair is made properly, and X' when repair is incorrect; the intermingled sequence of X and X' quantities are conditionally independent. Furthermore, let the probability of a correct repair at failure n be α, $0 \leq \alpha \leq 1$, if the repair at the time of previous failure $n-1$ was correct, and $\bar{\beta} = 1-\beta$ $0 \leq \beta \leq 1$ $(\alpha \neq \beta)$ if it was incorrect; the sequence of correct and incorrect repairs is thus modeled as a stationary ergodic Markov chain. This, of course, does not represent systematic improvements in repair capability, although a transient chain could serve for that purpose.

Let $M(t)$ $(M'(t))$ denote the mean or expected number of repairs in t, given that the first repair was correct (incorrect); think of the first repair (manufacture) as occurring at $t = 0$. Argue that

$$M(t) = \begin{cases} 1 & \text{if } X > t; \\ 1 + M(t-X) & \text{if } X \leq t \text{ and the 2nd repair is correct;} \\ 1 + M'(t-X) & \text{if } X \leq t \text{ and the 2nd repair is incorrect.} \end{cases} \quad (8.3)$$

Likewise,

$$M'(t) = \begin{cases} 1 & \text{if } X' > t; \\ 1 + M(t-X') & \text{if } X' \le t \text{ and the 2nd repair is correct;} \\ 1 + M'(t-X') & \text{if } X' \le t \text{ and the 2nd repair is incorrect.} \end{cases} \quad (8.4)$$

Now if the various conditions are removed then according to the model there results the two linked convolution integral equations:

$$M(t) = 1 + \alpha \int_0^t M(t-x) F_X(dx) + (1-\alpha) \int_0^t M(t-x) F_{X'}(dx). \quad (8.5)$$

and

$$M'(t) = 1 + \beta \int M(t-x) F_{X'}(dx) + (1-\beta) \int_0^t M(t-x) F_{X'}(dx). \quad (8.6)$$

In turn, these two equations are susceptible to transforming: multiply by e^{-st} and integrate to get

$$\hat{M}(s) = \frac{1}{s} + \alpha \hat{M}(s) \hat{F}_X(s) + \bar{\alpha} \hat{M}'(s) \hat{F}_X(s)$$

$$\hat{M}'(s) = \frac{1}{s} + \beta \hat{M}(s) \hat{F}_{X'}(s) + \bar{\beta} \hat{M}(s) \hat{F}_{X'}(s);$$

matrix notation is natural here, especially if more than two repair states are used. If one then solves and collects terms to order $(1/s)^2$ as $s \to 0$, Tauberian theorems, cf. Feller [1966] show that for large t

$$M(t) \approx M'(t) \sim \frac{\bar{\alpha} + \bar{\beta}}{\bar{\alpha} E[X'] + \bar{\beta} E[X]} t, \quad (8.7)$$

and a little reflection shows that this is entirely sensible. In similar ways variances can be written down, and an approximate Normal/Gaussian distribution for total failures may be derived.

The model can also be extended to account for the existence of non-zero repair times, and total availability studied.

Here is another, somewhat more complicated, application of renewal theory ideas, now to an inspected system problem.

Situation. (Standby system availability). A system, such as an emergency electric power source, is usually in a quiescent or cold standby status, but occasionally is called upon to fill a function (e.g. generate power). Various systematic plans might be devised for assuring reasonably high system availability, or probability of satisfying a demand when one occurs. One such plan is to inspect infrequently so long as no failure is detected between inspections, and otherwise inspect more frequently until evidence of need seems gone. Problem: develop a model to evaluate such an inspection scheme.

Model. Let the inspection plan be to inspect at long intervals, $\{L_i, i=1,2,...\}$ until such time as an inspection reveals a failure, and then switch to short intervals $\{S_i, i=1,2,...\}$, continuing until there has been a run of r(10,say) failure-free short-interval inspections, at which time switch back to long intervals; continue indefinitely. A measure of effectiveness is the long-run point availability of the system, i.e. the probability that the system is failure-free on the occasion of a demand.

To evaluate such a rule, allow the L_i and S_i sequences to be IID and mutually independent, with dfs $F_L(x)$ and $F_S(x)$ respectively; if desired these latter can be specialized to concentrate at fixed values (e.g. 14 days and 1 day, respectively.

Furthermore, let λ be the failure rate of a system failing at exponentially distributed intervals even when "cold." In order to analyze the system by renewal theory it is worthwhile to look at the periods during which inspection are infrequent, called L-eras, and those alternating with them, during which inspections occur frequently, called S-eras. Note that a demand can occur during either type of era, L-, and S-eras constitute an alternating renewal process.

The analysis involves the following components.

(A) Distribution (density) of L-era duration.

Suppose an inspection has just been completed at $t = 0$ and nothing amiss has been detected. Furthermore, suppose that this inspection marked the end of the previous S-era, so an L-era is just beginning. Let $a_L(dt)$ be the probability that the present L-era will last for time $(t,t+dt)$, or, loosely, until exactly t. One can now write down a renewal equation for $a_L(dt)$:

$$a_L(dt) = (1-e^{-\lambda t})F_L(dt) + \int_0^t e^{-\lambda t'}F_L(dt')a_L(dt-t'); \qquad (8.8)$$

the first term on the rhs means that the L-era terminates with the first inspection, meaning that the unit has failed before L_1. The second term represents survival through the first inspection at which time the process renews itself or starts anew; final failure occurs at time $t - t'$ thereafter. Failure or no failure at first inspection are mutually exclusive and exhaustive events, and so the result is a renewal equation for $a_L(dt)$. Introduce transforms to find

$$a_L(s) \equiv E[e^{-s\underline{L}}] = \frac{\hat{F}_L(s) - \hat{F}_L(\lambda+s)}{1 - \hat{F}_L(\lambda+s)} . \qquad (8.9)$$

The mean of the L-era duration, denoted by $\underline{L}$, is

$$E[\underline{L}] = \frac{E[L]}{1 - \hat{F}_L(\lambda)} . \qquad (8.10)$$

(B) Distribution of S-era duration.

If an inspection has just revealed a failure, an S-era begins immediately (take inspection and repair to be instantaneous for the moment). Let $a_S(dt)$ be the probability that an S-era lasts for time t. Then the following renewal equation may be written:

$$a_S(dt) = e^{-\lambda t} F_S^{r*}(dt) + \int_0^t h(dt')a_S(dt-t'); \qquad (8.11)$$

the auxiliary function h represents the probability that an inspection reveals a failure before the termination of the S-era in progress; this causes the frequent inspection to start over, i.e. starts the S-era afresh. In terms of transforms, $\underline{S}$ denoting an S-era duration,

$$\hat{a}_S(s) \equiv E[e^{-s\underline{S}}] = \frac{(\hat{F}_S(s+\lambda))^r}{1 - \hat{h}(s)} , \qquad (8.12)$$

and

$$\hat{h}(s) = [\hat{F}_S(s)-\hat{F}_S(\lambda+s)]\cdot\{1+\hat{F}_S(s+\lambda)+(\hat{F}_S(s+\lambda))^2+\ldots+(\hat{F}_S(s+\lambda))^{r-1}\}$$

$$= \frac{\hat{F}_S(s)-\hat{F}_S(s+\lambda)}{1-\hat{F}_S(s+\lambda)}\cdot[1-(\hat{F}_S(s+\lambda))^r] \quad ; \tag{8.13}$$

The latter transform expresses the probability that the necessary run of r successes is interrupted by a failure, and must begin again. From the transform comes the expected length of an S-era:

$$E[\underline{S}] = E[S]\{\frac{(\hat{F}_S(\lambda))^{-r}-1}{1-\hat{F}_S(\lambda)}\} \; . \tag{8.14}$$

(C) Probability that system is available during an L-era.

The probability $A_L(t)$ that the system is up at time t after the beginning of an L-era is simply $e^{-\lambda t}$, the probability of no failure, and the transform is

$$\int_0^\infty e^{-st} A_L(t)dt \equiv \hat{A}_L(s) = (\lambda+s)^{-1} \; . \tag{8.15}$$

(D) Probability that system is available during an S-era.

If $A_S(t)$ is the probability that the system is up after an S-era has progressed for time t, by renewal

$$A_S(t) = e^{-\lambda t}\bar{F}_S^{r*}(t) + \int_0^t h(dt')A_S(t-t') \tag{8.16}$$

where h has appeared before, under (B). It follows that

$$\hat{A}_S(s) = \frac{1 - (\hat{F}_S(s+\lambda))^r}{(s+\lambda)[1-\hat{h}(s)]} \quad . \tag{8.17}$$

(E) Overall availability at t.

Let $A(t)$ denote the availability of the system at t. Again by backward renewal argument, and starting at the beginning of an L-era,

$$A(t) = A_L(t) + \int_0^t A_S(t-t')a_L(dt') + \int_0^t A(t-t')a_L * a_S(dt') \; ; \tag{8.18}$$

transforms then give

$$\hat{A}(s) = \frac{\hat{A}_L(s) + \hat{A}_S(s)\hat{a}_L(s)}{1 - \hat{a}_L(s)\hat{a}_S(s)} \; ; \tag{8.19}$$

a Tauberian theorem now shows that

$$\lim_{t\to\infty} A(t) = \frac{\hat{A}_L(0) + \hat{A}_S(0)}{E[\underline{L}] + E[\underline{S}]} \; ,$$

so long-run point availability is

$$A(\infty) = (\lambda^{-1}) \left\{ \frac{(\hat{F}_S(\lambda))^{-r}}{E[L](1-\hat{F}_L(\lambda))^{-1} + E[S]((\hat{F}_S(\lambda))^{-r}-1)(1-\hat{F}_S(\lambda))^{-1}} \right\} \quad . \tag{8.20}$$

The expression for $A(\infty)$ is easily evaluated numerically if expressions for L-interval and S-interval transforms are obtainable. It is sometimes useful to interpret

$$\hat{A}(s) \cdot s = \int_0^\infty A(t)e^{-st}\, sdt \qquad (8.21)$$

as the availability of the system upon demand, the demand now occurring at a random time D, D having the exponential distribution with mean s^{-1}. In this case the initial conditions matter (they do not, in the long run), and a different availability figure is obtained depending upon whether the system is initially in an L-era or an S-era.

4. Additional Modeling Topics

In this section are outlines of certain modeling topics that formed the basis for cooperative research at ULB.

(1) Distributional Sculpturing or Inverse Modification

Standard distributions such as the Exponential or Gamma in particular, and also the Normal, log-Normal, and many others may reasonably and conveniently serve as components of a stochastic model or be used to summarize data distributions. For instance, inter-arrival times to service systems may appear approximately Exponential, service times nearly Gamma or log-Normal, and so on. On the other hand, a systematic departure from such a standard may be revealed by an analyzing actual data. It is then frequently possible to alter conventional distributions, or, equivalently, transform the random variable in simple ways in order to represent empirical reality more closely. Here are two conventional examples; there then follow some more general procedures.

Example 1. The Weibull distribution is often utilized to represent times to failure of components or times between system demands. Let T be a random variable (time, for instance), then T is distributed in a Weibull manner if

$$F_T(t;\alpha,\beta) = P\{T \leq t\} = \begin{cases} 1 - e^{-\alpha t^{\beta}} , & t > 0,\ \alpha > 0,\ \beta > 0 \\ 0 & t < 0 ; \end{cases} \tag{1.1}$$

The Weibull density is

$$f_T(t;\alpha,\beta) = e^{-\alpha t^{\beta}}\ \alpha\ \beta\ t^{\beta-1} . \tag{1.2}$$

Furthermore, the Weibull hazard or failure rate at age t is

$$\frac{1}{dt} P\{T \in (dt) \mid T > t\} \equiv k_T(t;\alpha,\beta)$$

$$= \frac{f_T(t;\alpha,\beta)}{1 - F_T(t;\alpha,\beta)} = \alpha\beta t^{\beta-1} ; \qquad (1.3)$$

The latter formula and also (1.2) imply that if $\beta = 1$ the Weibull is actually an Exponential; for this case "age" or "time in service" as measured by t does not influence the probability of failure in the next short time interval $(t,t+dt)$. On the other hand $\beta > 1$ implies that the rate of failure having survived to t--the hazard--increases with age, while, $\beta < 1$ implies that hazard decreases with age. If $\beta < 1$ the Weibull right tail is longer than that of the Exponential (extremely positively skewed), while if $\beta > 1$ the positive skewness is less pronounced. The Weibull r.v., T, is actually only an Exponential r.v., X, transformed or disguised, for if

$$P\{T \leq t\} = 1 - e^{-\alpha t^{\beta}} = P\{T^{\beta} \leq t^{\beta}\} , \qquad (1.4)$$

then

$$P\{T^{\beta} \leq x\} = 1 - e^{-\alpha x} .$$

Consequently, $T^{\beta} = X$, an Exponential r.v., and $T = X^{1/\beta}$ is a representation of a Weibull in terms of an Exponential. Supposing that one wishes to simulate a Weibull (α,β) r.v., then one merely simulates an Expon (1) r.v. and raises it to the

$(1/\beta)^{th}$ power, later multiplying by α^{-1} . Since the power transformation is monotonic increasing, the quantiles of Weibull and Exponential are related by

$$t(p) = (x(p))^{\beta} \tag{1.5}$$

Example 2. The log-Normal distribution is a favorite model for system repair times (see Kline and Almog [1980]; it has many other uses, and even some rational persuasions for its apparent resemblance to empirical distributions (see Aitchison and Brown [1957]). Say that X is log-Normal if $\ell n\ x = Y$ is Normal (it doesn't matter what the base of the logs is!); specifically $Y \sim N(\mu,\sigma^2)$. Here are some properties:

$$\cdot\quad m_k = e^{j\mu+\frac{1}{2}j^2\sigma^2} = E[X^k]$$

$$\cdot\quad m_1 = E[X] = e^{\mu+\frac{1}{2}\sigma^2}\ ,\ Var[X] = (E[X])^2(e^{\sigma^2}-1) \equiv (E[X])^2\eta^2$$

$$\cdot\quad \gamma_1(X) = skew[X] = \eta^3 + 3\eta \tag{1.6}$$

$$\cdot\quad \gamma_2(X) = kur[X] = \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2$$

$$\cdot\quad Median[X] = e^{\mu}\ ,\ Mode[X] = e^{\mu-\sigma^2}$$

Supposing that one wishes to simulate a log-Normal random variable, one simply simulates a $N(\mu,\sigma^2)$ r.v., Y, and exponentiates $X = e^Y$.

The above two familiar examples illustrate the formation of new and useful random variables and distributions by simple transformation. An intuitively appealing way of looking at certain transformations is in terms of modifications to familiar random variables or quantiles by convenient shaping factors. This process will be called

Distributional Sculpturing: Let X be a basic r.v., for example an Exponential, Normal or log-Normal. Define

$$Y = Xs(X) \tag{1.7}$$

where the shaping function $s(X)$ is designed to conveniently convert the basic r.v. X into a shaped version, Y, having desired distributional properties. Some examples of the properties often desired in practice, along with suitable--but not unique--shaping functions, now follow.

(A) Skewness-Producing Shapers

Example: X is a positive basic r.v., e.g., Exponential.

$$\text{(i)} \quad s(X) = 1 + AX^{\ell}, \quad A > 0, \quad \ell > 0 \tag{1.8}$$

$$\text{(ii)} \quad s(X) = e^{AX^{\ell}}$$

Effect:

$$Y = X\,s(X) \simeq \begin{cases} X & \text{if } X \text{ "small"} \\ \text{(i)} \quad X + AX^{\ell+1} \gg X & \\ \text{(ii)} \quad Xe^{AX^{\ell}} & \end{cases} \quad \begin{array}{l} \text{much greater} \\ \text{than } X \text{ if } X \\ \text{"large"} \end{array}$$

The particular shaping functions (i) and (ii) both leave small values of X unchanged, but considerably expand large values, thus transforming the distribution of X, e.g., the Exponential, to one that is nearly Exponential near the origin, but having a relatively long right tail. For example, take shaping function (2) with $\ell = 1$ and apply to X Exponential. The shaped distribution becomes

$$F_Y(y;A) = 1 - \exp[-(\frac{\sqrt{1+4Ay} - 1}{2A})] \ . \tag{1.9}$$

Examination shows that for small y (Taylor series) the distribution of Y is nearly unit Exponential, while for large Y it resembles a Weibull with shape parameter $\beta = 1/2$. Shaping function (ii) has an even more pronounced effect on the right tail. Note that both shaping functions (i) and (ii) yield monotonic increasing transformations from $X \to Y$, and that given by (i) is sometimes explicitly algebraically invertible (solve quadratic equation when $\ell = 1$, cubic when $\ell = 2$, quartic when $\ell = 3$), while that of (ii) is not. Also note that useful transformations result when parameter $A < 0$: this actually may result in right tail truncation (severe shortening; such transformations are no longer monotonic.

• Moments.

The moments of shaped or sculptured r.v.s. can sometimes be conveniently calculated. For the present representations:

(i) $Y = X(1 + AX^{\ell})$;

$$m_1(Y) = m_1(X) + Am_{1+\ell}(X)$$

$$m_2(Y) = E[(X(1 + AX^{\ell}))^2]$$

$$= m_2(X^2) + 2Am_{1+\ell}(X) + A^2 m_{2+2\ell}(X) \qquad (1.10)$$

$$Var[Y] = Var[X] + 2ACov[X, X^{\ell+1}] + A^2 Var[X^{\ell+1}] .$$

Also, to indicate the dependence between the stretched r.v. Y and the basic r.v. X,

$$m_k(Y-X) = A^k m_1(X^{(\ell+1)k}) = A^k m_{(\ell+1)k}(X) ,$$

so

$$Var[Y-X] = A^2 Var[X^{\ell+1}] \qquad (1.11)$$

and

$$Cov[Y,X] = Var[X] + ACov[X, X^{\ell+1}] . \qquad (1.12)$$

The quantiles are directly and simply related by

$$y(p) = x(p)(1 + A(x(p))^{\ell}) . \quad 0 \leq p \leq 1 \qquad (1.13)$$

(ii) $Y = Xe^{AX^{\ell}}$; but for the present consider only $\ell = 1$. Then the k^{th} moment is expressible in terms of the derivatives of the Laplace transform of X . Note that the k^{th} moment does not necessarily exist for all basic distributions.

When it does,

$$m_k(Y) = (-1)^k \frac{d}{ds} E[e^{-sX}]\Big|_{s = -kA} \qquad (1.14)$$

For example, if $X \sim$ Expon (1),

$$m_k(Y) = \frac{k!}{(1 - kA)^{k+1}} \quad \text{if} \quad kA < 1; \qquad (1.15)$$

otherwise the moment doesn't exist because the distribution's right tail is too long.

(B) Symmetric Stretch-Producing Shapers

Example: Z is a r.v. symmetrically distributed around zero; e.g., $Z \sim N(0,\sigma^2)$. Here are some useful shapers:

(i) $s(Z) = 1 + hZ^2$, $h > 0$;

(ii) $s(Z) = e^{hz^2}$ (due to J. W. Tukey).

Again (i) and (ii) imply that $Y = Zs(Z)$ resembles Z for small Z, but lengthens the tails of the distribution for large Z .

Example 1. Stretched log-Normal variables may be suggested for modeling repair or service times if data analysis indicates that the logarithms of observwd times are symmetric but not nearly Normal, having symmetrically too-long tails. Then it may be convenient to use the representation

$$Y = e^X$$

$$X = \mu + cZs(Z) \tag{1.16}$$

where μ is the center (corresponds to mean) of the logged observations, and the sale constant c is the standard deviation (spread parameter) of the variable $Zs(Z)$, replacing σ in the ordinary log-Normal formula.

• Moments.

The moments of the shaped distribution are, for (i), representable in terms of those for a basic Z.

(i) $Y = Z(1 + hZ^2)$

$$\begin{aligned} m_1(Y) &= m_1(Z) + hm_3(Z) , \\ &= 0 \text{ (Z symmetrical around zero).} \end{aligned} \tag{1.17}$$

$$m_2(Y) = m_2(Z) + 2hm_4(Z) + h^2m_6(Z) .$$

$$m_4(Y) = m_4(Z) + 4hm_6(Z) + 6h^2m_8(Z) + 4h^3m_{10}(Z) + h^4m_{12}(Z) .$$

(ii) $Y = Ze^{hZ^2}$; consider only $Z \sim N(0,\sigma^2)$. Calculate as a preliminary

$$\int_{-\infty}^{\infty} e^{hz^2} e^{-\frac{1}{2}z^2/\sigma^2} \frac{1}{\sqrt{2\pi}\sigma} = (1 - 2h\sigma^2)^{-1/2} , \quad h < 2\sigma^2 \tag{1.18}$$

Then differentiate repeatedly with respect to h to obtain the even moments (the odd moments equal zero):

$$\text{Var}[Y] = m_2(Y) = \frac{\sigma^2}{(1 - 4h\sigma^2)^{3/2}}, \qquad h < 1/4\sigma^2 ;$$

$$m_4(Y) = \frac{3\sigma^4}{(1 - 8h\sigma^2)^{5/2}}, \qquad h < 1/8\sigma^2 . \tag{1.19}$$

Hence kurtosis is

$$\gamma_2 = \frac{m_4(Y)}{(m_2(Y))^2} = \frac{3(1 - 4h\sigma^2)^3}{(1 - 8h\sigma^2)^{5/2}} - 3 \tag{1.20}$$

Tails are so extended that the kurtosis becomes very large in the case of (i), and actually infinite for rather small values of h in (ii); the variance remains finite for slightly larger values. Nevertheless, the central part of the Y-distribution remains remarkably close to the Normal from when Z is itself Normal.

Both forms (i) and (ii) can be induced to fit the inverse distribution (percent points) of the Student's t distribution fairly satisfactorily.

(C) Left Tail Enhancement Shapers

Example: X is a positive r.v., e.g. Exponential.

(i) $s(X) = \dfrac{\alpha X^{\ell}}{1 + \alpha X^{\ell}}, \qquad \ell > 0, \ \alpha > 0;$

$$\text{(ii)} \quad s(X) = 1 - e^{-\alpha X^{\ell}} \; ;$$

$$\text{(iii)} \quad s(X) = e^{-\alpha / X^{\ell}} \; . \tag{1.21}$$

This shaping function tends to concentrate probability near zero, leaving the distribution's shape unchanged for large X :

$$Y = Xs(X) \simeq \begin{cases} X & \text{if } X \text{ large} \\ \left.\begin{matrix} \text{(i)} \\ \text{(ii)} \end{matrix}\right\} & \alpha X^{\ell+1} << X, \; X \text{ small} \end{cases} \tag{1.22}$$

Moments are generally impossible to find in useable form, although in the case of (ii), $\ell = 1$, explicit results can be found in terms of Laplace transforms: suppose $X \sim \text{Expon}(1)$, then under (ii)

$$E[Y] = E[X(1 - e^{-\alpha X})] = 1 - \frac{1}{(1 + \alpha)^2} \; ; \tag{1.23}$$

the mean of Y is influenced very little when α is large, but small values of X are made even smaller; for the Expon(1) X the quantiles are related as follows:

$$y(p) = x(p)[1 - e^{\alpha \ln(1-p)}] \; , \tag{1.24}$$

so for a fixed $p > 0$ large α forces $y(p)/x(p)$ to one. However, for a fixed α and small p

$$\frac{y(p)}{x(p)} \simeq \alpha p \simeq 0 \ , \tag{1.25}$$

showing that low quantiles for X transform into even smaller values for Y. Furthermore, it is easy to show that if $X_{(1;n)}$ is the minimum in a sample of n from $X \sim \text{Expon}(1)$, and

$$Y_{(1;n)} = X_{(1;n)}[1 - e^{-\alpha X_{(1;n)}}] \ ,$$

then

$$\frac{E[Y_{(1;n)}]}{E[X_{(1;n)}]} = \frac{\alpha}{n + \alpha} \to 0 \quad \text{as} \quad n \to \infty \ . \tag{1.26}$$

All of this reinforces the image of the present shaper as forcing small X-values to become smaller, leaving large values unchanged. Note that the result of the shaper (i), $\ell = 1$, can be explicitly inverted, giving the distribution

$$F_Y(y) = F_X\left[\frac{y(1 + \sqrt{1+4/\ y})}{2}\right] \ , \tag{1.27}$$

from which the left-tail enhancement property shows itself explicitly:

$$f_Y(y) = f_X\left|\frac{y(1+\sqrt{1+4/\alpha y})}{2}\right| \cdot \left\{1 + \frac{y + (4/\alpha)}{y^2 + (\frac{4}{\alpha})y}\right\} \cdot (\tfrac{1}{2})$$

(1.28)

$$\simeq \begin{cases} f_X(\frac{\sqrt{y}}{\sqrt{\alpha}}) \frac{1}{\sqrt{\alpha}\sqrt{y}} & \text{for } y \text{ small} \\ f_X(y) & \text{for } y \text{ large} \end{cases}$$

Note that the hazard associated with the latter transformation is, when X is an Expon(1) r.v.,

$$h_Y(y) = \tfrac{1}{2}\{1 + \frac{y + 4/\alpha}{\sqrt{y^2 + (4/\alpha)y}}\}$$

$$\simeq \begin{cases} 1/\sqrt{\alpha y} & \text{for } y \text{ small} \\ 1 & \text{for } y \text{ large} \end{cases} \quad (1.29)$$

Thus the density of an Exponential X remains, under this transformation, Exponential-like in the right tail but approaches infinity near the origin. A distribution with such behavior may well be useful for modeling failure data exhibiting early failures ("infant mortality.")

(D) Right-Tail - Shortening Shaper

Example: X is a positive r.v., e.g. an Exponential. To shorten or truncate the right tail, consider

$$\text{(i)} \quad s(X) = \frac{1}{1 + \beta X^{\ell}}, \qquad \beta > 0, \quad \ell > 0$$

$$\text{(ii)} \quad s(X) = e^{-\alpha X^{\ell}}, \qquad \alpha > 0, \quad \ell > 0 \tag{1.30}$$

The shaper (i) with $\ell = 1$ is a right tail shortener or truncator; it provides a monotonic transformation to $Y \leq \beta^{-1}$; small values of X are left nearly unchanged. The quantiles of Y are

$$y(p) = \frac{x(p)}{1 + \beta x(p)} \simeq \begin{cases} x(p) & \text{for } p \text{ small} \\ \\ \beta^{-1} & \text{for } p \text{ large if } x(p) >> \beta^{-1} \end{cases} \tag{1.31}$$

Shaper (ii) is non-monotonic: values of $Y = Xs(X)$ increase to $(\alpha\ell)^{-1}$, decreasing thereafter. The inverse of (i) yields

$$F_Y(y) = F_X\left(\frac{y}{1 - \beta y}\right), \qquad 0 \leq y \leq \beta^{-1} \tag{1.32}$$

with density

$$f_Y(y) = f_X\left(\frac{y}{1 - \beta y}\right) \frac{1}{(1 - \beta y)^2}; \tag{1.33}$$

the latter approaches infinity at a rapid rate as y nears $1/\beta$.

Shaper (i) with $\ell = 1/2$ is a tail thinner; it can be inverted to give

$$F_Y(y) = F_X[(\beta y)^2(\frac{1 + \sqrt{1+(4/\beta y)}}{2})^2] \; ; \tag{1.34}$$

with density

$$f_Y(y) = f_X[\beta y)^2(\frac{1 + \sqrt{1+(4/\beta y)}}{2})^2][\frac{\beta^2 y}{2}(1+\sqrt{1+(4/\beta^2 y)})(1 + \frac{1+(2/\beta y)}{\sqrt{1+(4/\beta^2 y}})]$$

$$\simeq \begin{cases} f_X(y) & \text{for } y \text{ small} \\ \\ f_X(\beta^2 y^2) 2\beta^2 y & \text{for } y \text{ large} \; ; \end{cases} \tag{1.35}$$

if X is Exponential, then the hazard rate of Y is initially flat, but eventually increases linearly with age, thus producing a plausible wearout model for equipment or biological organisms.

The above examples illustrate a few of the large number of possible ways in which the sculpturing idea can be used to extend the descriptive power of standard distributional models.

Problems of fitting such models to actual data are currently being addressed, as are applications to simulation and time series modeling. Use of shaped exponentials to evaluate scheduling procedures has been initiated in collaboration with P. Collard at ULB.

(2) Response Times Under Processor Sharing

Consider a simple model of a time-sharing computer system in which N terminals access a single computer (server). Let repairman-model conditions prevail, but processor sharing governs the service order: if j jobs ($0 \leq j \leq N$) are at the execution stage each receives one-j^{th} of a time unit of service per time unit. In other words, if it is the service rate (μ^{-1} is the expected service time under Markov-exponential assumptions), then the unconditional probability that any designated single job finishes in $(t, t + s)$ is $\mu(\frac{\Delta}{j}) + 0(\Delta)$.

Under such conditions it is possible to derive backward-type equations to describe the response of waiting time, R , of a newly-initiated "tagged" job arriving from a previously idle terminal. In particular, consider

$$m_j(t) = E[R|W(R) = t \ , \ X(0) = j] \ , \tag{2.1}$$

where $W(R)$ is the amount of work or processing time required of the server, and $X(0)$ represents the number of jobs currently at the processor when the tagged job first arrives (including that job). Thus $m_j(t)$ is the expected response time, conditional on need and accompaniment. Additionally, introduce $r(j)$ as the fraction of a time quantum, D , actually available for job processing when j jobs are present at the computer; $r(j)$ represents one component of overhead, and may decrease as j increases. For short let λ_j represent the rate of new arrivals when j are present at the server;

under stated model conditions $\lambda_j = \lambda(N-j)$. Let μ_j be the rate at which jobs accompanying the tagged job depart; under stated conditions $\mu_j = \mu(j-1)\left(\frac{r(j)\Delta}{j}\right) + 0(\Delta)$. These assumptions, or those for a general birth and death process, lead by a backward-conditioning to this equation:

$$m_j(t) = \Delta + m_j\left(t - \left(\frac{r(j)}{j}\right)\Delta\right)\left[1 - (\lambda_j + \mu_j)\Delta\right] + \lambda_j \Delta m_{j+1}\left(t - \frac{r(j)\Delta}{j}\right)$$

$$+ \mu_j \Delta m_{j-1}\left(t - \frac{r(j)\Delta}{j}\right) + o(\Delta) \ . \qquad (2.2)$$

Subtract $m_j(t - \frac{r(j)\Delta}{j})$ from each side, divide by Δ , and let $\Delta \to 0$ to obtain

$$\left(\frac{r(j)}{j}\right)\frac{dm_j(t)}{dt} = 1 - (\lambda_j + \mu_j)\, m_j(+) + \lambda_j\, m_{j+1}(t) + \mu_j\, m_{j-1}(+) \qquad (2.3)$$

This is a standard system of linear differential equations with constant coefficients that may be solved by standard methods.

If $r(j)$ is constant, and the repairman model assumptions are fulfilled, then it has been shown by G. Latouche that

$$E[R|W(R) = t] = C\, t \ , \qquad (2.4)$$

i.e. is linear in t , with C depending upon λ and μ . See the article by Mitra [1981] for more detail concerning this problem.

(3) Repairman Model in Random Environment

Suppose we have m machines (this may mean electric power generators, or even remote computer terminals) that, when in use, fail independently (computer terminals: apply for processing time, or data) at rate $\lambda(t)$, and, if failed are repaired at rate $\mu(t)$; all processes are Markovian, given $\lambda(t)$ and $\mu(t)$, $t \geq 0$. Now let $\lambda(t)$ and $\mu(t)$ themselves be realizations of a finite-state Markov process that develops independently of the number $N(t)$ of machines down for repair. If $N(t)$ is the number down at t, and $J(t)$ is the underlying environmental state, then $\{N(t), J(t)\}$ is a bivariate Markov process, and $N(t)$ change is governed by the current level of the environment $J(t)$. The latter environment may refer to physical conditions such as heat, seismic shock, or to variations in repair effectiveness. In the case of computer terminals or communication nodes the environmental variations may be the result of changes in message transmissions or data demands under occasional crisis conditions.

The paper by Gaver, Jacobs, and Latouche [1981] presents a systematic mathematical analysis of the general birth-and-death process in random environments, including the above repairman model as a special case. Numerical illustrations are provided. Here we present a truncated version of the solution to the first-passage time problems, utilizing a recursive or "clawing-up" mode of thinking analogous to developments in Section ( ) of this account. Restrict discussion to just two

environmental levels, denoted by $j = 1, 2$. Let the transition rate from environmental state $j = 2 \to 1$ be α , and from $j = 1 \to 2$ be β . Let $\lambda_n(j)$ be the transition rate from $(n,j) \to (n+1, j)$, and $\mu_n(j)$ be that from $(n,j) \to (n-1, j)$.

Let, as before, U_n be the first-passage time from n to $n + 1$, and put

$$G_n(dx;i,j) = P\{U_n \varepsilon (dx), J(U_n) = j | N(0) = n, J(0) = i\} , \qquad (3.1)$$

with $i, j = 1, 2$; the L.-S. transform of this measure is called

$$G_n(i,j;s) = E[e^{-sU_n}, J(U_n) = j | N(0) = n, J(0) = i]; \qquad (3.2)$$

this is the transform of the time to pass for the first time from a state in which n are down, the environment being in state i at some initial instant $(t = 0)$, to $n + 1$ down, the environment being in state j . Simple considerations of cases that may arise during the first transition give, first for $i = 1$,

$$G_n(1,j;s) = \frac{\lambda_n(1)}{d_n(1)} \ell_j(1) + \frac{\mu_n(1)}{d_n(1)} \sum_{k=1}^{2} G_{n-1}(1,k;s)\, G_n(k,j;s) + \frac{\alpha}{d_n(1)} G_n(2,j;s) \qquad (3.3)$$

where $j = 1, 2$, the indicator function

$$\ell_i(j) = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

and the denominator

$$d_n(i) = \lambda_n(1) + \mu_n(1) + \beta + s \ . \tag{3.5}$$

Likewise, for $i = 2$,

$$G_n(2,j;s) = \frac{\lambda_n(2)}{d_n(2)} \ell_j(2) + \frac{\mu_n(2)}{d_n(2)} \sum_{k=1}^{2} G_{n-1}(2,k;s)\, G_n(k,j;s) + \frac{\beta}{d_n(2)} G_n(1,j;s) \tag{3.6}$$

(the equivalent of these equations in Chu and Gaver [ ] accidentally incorrectly omits the final term on the rhs). The above equations can in principle be solved recursively, beginning with

$$G_0(1,j;s) = \frac{\lambda_0(1)}{d_0(1)} \ell_j(1) + \frac{\alpha}{d_0(1)} G_0(2,j;s)$$

$$G_0(2,j;s) = \frac{\lambda_0(2)}{d_0(2)} \ell_j(2) + \frac{\beta}{d_0(2)} G_0(1,j;s) \qquad j = 1,2. \tag{3.7}$$

where

$$d_0(1) = \lambda_0(1) + \beta + s \ . \qquad d_0(2) = \lambda_0(2) + \alpha + s \ .$$

The first-passage time T_n from $N(0) = 0$ and $J(0) = i$ $(i=1,2)$ to $n + 1$ has transform

$$P_n(i,j;s) = E[e^{-sT_n}, J(T_n) = j | N(0) = 0, J(0) = i] ;$$

in matrix notation,

$$\mathcal{P}_n = \mathcal{G}_0 \mathcal{G}_1 \cdots \mathcal{G}_n$$

and the L.-S. transform of the first-passage time to $n + 1$ in $\mathcal{P}_n \mathcal{l}$, where $\mathcal{l} = (1,1)^T$, a column vector.

Differentiation of expressions (3.2) and (3.5) produces recursive expressions for means, variances and higher moments. Programs have been written to evaluate these expressions numerically.

REFERENCES

[1] AITCHISON J., and BROWN J.A.C. (1957). The Log-Normal Distribution. Cambridge University Press, Cambridge, England.

[2] BENDER E.A. (1978). An Introduction to Mathematical Modeling. John Wiley and Sons, Inc., New York.

[3] BURMAN D.Y. (1979). An Analytic Approach to Diffusion Approximations in Queueing. Ph.D. Dissertation, NYU Courant Inst.

[4] COX D.R., and SMITH, W.L. (1961). Queues. Methuen Monograph. John Wiley and Sons, Inc., New York.

[5] COX D.R. (1962). Renewal Theory. Methuen Monograph. John Wiley and Sons, Inc., New York.

[6] COX D.R. (1969). Analysis of Binary Data. Chapman and Hall. London.

[7] CHU B.B., and GAVER D.P. (1977). Stochastic models for repairable redundant systems susceptible to common mode failure. S.I.A.M., Proc. of Int. Conf. on Nuclear Systems, Reliab. Eng. and Risk Assessment, Gatlinburg, Tenn., 342-367.

[8] CHUNG K.L. (1967). Markov Chains with Stationary Transition Probabilities. 2nd Ed. Springer-Verlag, New York.

[9] CRAMER H. (1946). Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ.

[10] DANIELS H.E. (1954). "Saddlepoint approximations in statistics". Ann. of Math. Statistics, Vol. 25, pp. 631-650.

[11] FELLER W. (1951, 1966). An Introduction to Probability Theory and Its Applications Vols. I and II. John Wiley and Sons, Inc., New York.

[12] GAVER D.P., and JACOB P.A. (1981). "On combinations of random loads". S.I.A.M. Jnl. Appl. Math., Vol. 40, pp. 454-466.

[13] GAVER D.P. and LEHOCZKY J.P. (1981). "Diffusion approximations for the cooperative service of voice and data messages". J. Appl. Prob., Vol. 18, pp. 660-671.

[14] GAVER D.P., JACOBS P.A., and LATOUCHE G. (1981). "Finite birth-and-death models in randomly changing environments". Report Interne No. 121, Lab. D'Informat. Theor., Univ. Libre de Brux.

[15] GAVER D.P., and THOMPSON G.L. (1973). Programming and Probability Models in Operations Research. Brooks-Cole Publ. Co., Monterey, CA.

[16] GAVER D.P., and SHEDLER G. (1973). "Approximate models for processor utilization in multiprogrammed computer systems". S.I.A.M. J. Comput., Vol. 2, No. 3, pp. 183-192.

[17] GAVER D.P. (1969). "Highway delays resulting from flow-stopping incidents". J. App. Prob., Vo. 6, pp. 137-153.

[18] GAVER D. P. (1962). "A waiting line with interrupted service, including priorities". J. of the Royal Stat. Soc., B, Vol. 25, pp. 73-90.

[19] HARRISON J.M., and REIMAN M.I. (1981). "Reflected Brownian motion on an orthant". Annals. Prob., Vol. 9, pp. 302-308.

[20] JACOBS P.A. (1980). "Heavy traffic results for single-server queues with dependent (EARMA) service and arrival times. Adv. Appl. Prob., Vol. 12, pp. 517-529.

[21] JACOBS P.A. (1978). "A cyclic queueing network with dependent exponential service times". J. of Appl. Prob., Vol. 15, pp. 573-589.

[22] KARLIN S., and TAYLOR H.M. (1975). A First Course in Stochastic Processes. Academic Press, London.

[23] KLEINROCK L. (1976). Queueing Systems, Vols. I, II. John Wiley and Sons, Inc., New York.

[24] KLINE, M.B., and ALMOG, R. [1980]. "Suitability of the loognormal distribution for repair times. Second Inst. Conf. on Reliability and Maintainability Perros-Guirec-Trégastel, France, Sept. 1980.

[25] KOBAYASHI H. (1981). This volume.

[26] LEVY D., and KAHN E. (1981). "Accuracy of the Edgeworth expansion of Loss-of-Load Probability calculations in small power systems". Summer Mtg., IEEE Power Eng. Soc. (IEEE Trans. on Power Apparatus and Systems, to appear).

[27] LEWIS P.A.W., and SHEDLER G.S. (1979). "Simulation of non-homogeneous Poisson processes by thinning". Naval Res. Log. Quart., Vol. 26, No. 3, pp. 403-413.

[28] LUGANNANI R., and RICE S. (1980). "Saddle point approximation for the distribution of the sum of independent random variables". Adv. in Applied Prob., Vol. 12, No. 2, pp. 475-490.

[29] McNEIL D.R., and SCHACH S. (1973). "Central limit analogues for Markov population processes" (with discussion). J. Roy. Stat. Soc., Vol. 35, pp. 1-23.

[30] McNEIL D.R. (1970). "Integral functionals of birth and death processes and related limiting distributions". Annals. of Statistics, Vo. 41, No. 2, pp. 480-485.

[31] MITRA D. (1981). "Waiting time distributions from closed queueing network models of shared-processor systems". Proc. Eighth Internat. Symp. Comp. Performance Modelling Measurem. and Eval. Amsterdam, The Netherlands.

[32] NEUTS M.F. (1981). Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach. The Johns Hopkins University Press, Baltimore Maryland.

[33] NEUTS M.F., and MEIER K.S. (1981). On the use of phase type distributions in reliability modelling of systems with a small number of components. OR Spektrum, 2, 227-234.

[34] NEWELL G.F. (1979). Approximate Behavior of Tandens Queues. Springer-Verlag. Berlin, Heidelber, New York.

[35] NEWELL G.F. (1971). Applications of Queueing Theory. Chapman and Hall. London, England.

[36] PREGIBON D. (1981). "Logistic regression diagnostics". Ann. of Statistics, Vol. 9, pp. 705-724.

[37] REIMAN M.I. (1982). Open queueing networks in heavy traffic". To appear in Math. of Opns. Res.

[38] TAKACS L. (1962). Introduction to the Theory of Queues Oxford University Press, New York, NY.

DISTRIBUTION LIST

	NO. OF COPIES
Library, Code 0142 Naval Postgraduate School Monterey, CA 93940	4
Dean of Research Code 012A Naval Postgraduate School Monterey, CA 93940	1
Library, Code 55 Naval Postgraduate School Monterey, CA 93940	2
Professor D. P. Gaver Code 55Gv Naval Postgraduate School Monterey, CA 93940	148
Chief of NavalResearch Arlington, Virginia 22217	2
Defense Technical Information Center ATTN: DTIC-DDR Cameron Station Alexandria, Virginia 22314	2